12

RIDP *libri*

Katalin Ligeti, John Vervaele, Gert Vermeulen
(Eds.)

Artificial Intelligence and Criminal Justice

(Concept paper for the 2020-2024 IAPL cycle and resolutions of the XXIst International Congress of Penal Law, Paris, 25-28 June 2024)

Intelligence Artificielle et Justice Pénale

(Résolutions du XXIe Congrès International de Droit Pénal, Paris, 25-28 juin 2024)

Inteligencia Artificial y Justicia Penal

(Resoluciones del XXI Congreso Internacional de Derecho Penal, París, 25-28 junio 2024)

Revue Internationale de Droit Pénal
International Review of Penal Law
Revista internacional de Derecho Penal
Международное обозрение уголовного права
刑事法律国际评论
المجلة الدولية للقانون الجنائي
Revista Internacional de Direito Penal
Rivista internazionale di diritto penale
Internationale Revue für Strafrecht

MAKLU

Artificial Intelligence and Criminal Justice

(Concept paper for the 2020-2024 IAPL cycle and resolutions of the XXIst International Congress of Penal Law, Paris, 25-28 June 2024)

Intelligence Artificielle et Justice Pénale

(Résolutions du XXIe Congrès International de Droit Pénal, Paris, 25-28 juin 2024)

Inteligencia Artificial y Justicia Penal

(Resoluciones del XXI Congreso Internacional de Derecho Penal, París, 25-28 junio 2024)

Artificial Intelligence and Criminal Justice
Intelligence Artificielle et Justice Pénale
Inteligencia Artificial y Justicia Penal

Edited by

Katalin Ligeti
John Vervaele
Gert Vermeulen

RIDP
Revue Internationale de Droit Pénal
International Review of Penal Law
Revista internacional de Derecho Penal
Международное обозрение уголовного права
国际刑事法律评论
المجلة الدولية للقانون الجنائي
Revista Internacional de Direito Penal
Rivista internazionale di diritto penale
Internationale Revue für Strafrecht

Maklu

Antwerpen | Apeldoorn | Portland
Maklu Publishers
Somersstraat 13/15, 2018 Antwerpen, Belgium, info@maklu.be
Koninginnelaan 96, 7315 EB Apeldoorn, The Netherlands, info@maklu.nl
www.maklu.eu

USA & Canada
International Specialized Book Services
920 NE 58th Ave., Suite 300, Portland, OR 97213-3786, orders@isbs.com, www.isbs.com

AIDP – Association Internationale de Droit Pénal | The International Association of Penal Law is the oldest association of specialists in penal law in the world. Since 1924, it is dedicated to the scientific study of criminal law and covers: (1) criminal policy and codification of penal law, (2) comparative criminal law, (3) international criminal law (incl. specialization in international criminal justice) and (4) human rights in the administration of criminal justice. The Association's website provides further information (http://www.penal.org).

RIDP – Revue Internationale de Droit Pénal | The International Review of Penal Law is the primary publication medium and core scientific output of the Association. It seeks to contribute to the development of ideas, knowledge, and practices in the field of penal sciences. Combining international and comparative perspectives, the RIDP covers criminal law theory and philosophy, general principles of criminal law, special criminal law, criminal procedure, and international criminal law. The RIDP is published twice a year. Typically, issues are linked to the Association's core scientific activities, ie the AIDP conferences, Young Penalist conferences, world conferences or, every five years, the International Congress of Penal Law. Occasionally, issues will be dedicated to a single, topical scientific theme, validated by the Scientific Committee of the Association, comprising high-quality papers which have been either presented and discussed in small-scale expert colloquia or selected following an open call for papers. The RIDP is published in English only.

RIDP *libri* | In conformity with the AIDP publication housestyle, the RIDP *libri* series allows book to be published under the RIDP AIDP flag. These mono-lingual books can be published in any of the six UN languages (Arab, Chinese, English, French, Russian, Spanish), German, Italian and Portuguese. This series offers great opportunity for scientific work to be shared across the organisation bringing together practitioners, academics and civil servants in different countries and regions across the world.

Peer review | All contributions are subject to double-layered peer review. The primary scientific and peer review responsibility for all issues lies with the designated Scientific Editor(s). The additional scientific quality control is carried out by the Executive Committee of the Editorial Board, which may turn to the Committee of Reviewers for supplementary peer review.

Disclaimer | The statements and opinions made in the RIDP *libri* contributions are solely those of the respective authors and not of the Association or MAKLU Publishers. Neither of them accepts legal responsibility or liability for any errors or omissions in the contributions nor makes any representation, express or implied, with respect to the accuracy of the material.

ISBN 978-90-466-1296-5
D/2025/1997/21
NUR 824
BISAC LAW026000
Theme: LNF, LAR

Summary

Preface
Katalin Ligeti, John Vervaele & Gert Vermeulen 7

Préface
Katalin Ligeti, John Vervaele & Gert Vermeulen 9

Prefacio
Katalin Ligeti, John Vervaele & Gert Vermeulen 11

Concept paper for the 2020-2024 IAPL cycle

Artificial intelligence and criminal justice
Katalin Ligeti 15

Resolutions / Résolutions / Resoluciones

Resolutions of the XXIst International Congress of Penal Law, Paris, 25-28 June 2024 35

Résolutions du XXIe Congrès International de Droit Pénal, Paris, 25-28 juin 2024 65

Resoluciones del XXI Congreso Internacional de Derecho Penal, París, 25-28 junio 2024 97

PREFACE

Katalin Ligeti, John Vervaele and Gert Vermeulen[*]

Criminal justice systems around the world are now routinely integrating computer technologies based on artificial intelligence (AI) into the detection, investigation, and prosecution of crime. AI's promise of rapid analysis and seemingly objective, science-based results makes it particularly attractive in the legal domain. There is a strong temptation to believe that AI can eliminate the "human factor," often perceived as emotional, fallible, and a potential threat to correct decision-making.

To harness these perceived advantages, AI developers have begun offering their services to judicial authorities and other actors in criminal proceedings. They create proprietary algorithms and statistical models tailored specifically for the prevention, repression, and adjudication of crime.

Recognizing the profound impact of these developments, the IAPL devoted its 2020-2024 scientific cycle and its concluding XXI[st] Congress to examining the transformative effects of AI and the legal challenges it poses for both substantive and procedural criminal law. The topic was prepared through a concept paper and explored through four international colloquia, each culminating in the adoption of a set of resolutions addressing a distinct aspect of the criminal justice system.

This volume presents the 2019 concept paper and the adopted resolutions, structured according to the four foundational sections of our Association:

- the Resolutions on "Traditional Criminal Law Categories and AI", adopted during the Section 1 colloquium on the general part of criminal law, held in Siracusa from 14-17 September 2022;
- the Resolutions on "Penal Law and Criminalization in the Face of the Challenges of AI", adopted during the Section 2 colloquium on the special part of criminal law, held in Bucharest from 14 to 16 June 2023;
- the Resolutions on "AI and the Administration of Criminal Justice: Predictive Policing, Predictive Justice, and Evidence", adopted during the Section 3 colloquium on criminal procedure, held in Buenos Aires from 28 to 31 March 2023;

[*] Katalin Ligeti is President of the AIDP/IAPL and Dean of the Faculty of Law, Economics and Finance and Professor of European and International Criminal Law at the University of Luxembourg. John A.E. Vervaele is Honorary President of the AIDP/AIPL, Emeritus Professor at Utrecht University, The Netherlands, and Professor in European Criminal Law and Human Rights at the College of Europe, Bruges, Belgium. Gert Vermeulen is General Director Publications of the AIDP/IAPL, Editor-in-chief of the RIDP, and Senior Full Professor of European and international criminal law, sexual criminal law, and data protection law at Ghent University, Belgium.

- the Resolutions on "International Perspectives on AI: Challenges for Judicial Co-operation and International Humanitarian/Criminal Law", adopted during the Section 4 colloquium on international criminal law, held in Opatija from 7 to 8 December 2023.

This significant scholarly achievement would not have been possible without the outstanding work of the four General Rapporteurs: Prof. Lorenzo Picotti (Section 1), Prof. Fernando Miró (Section 2), Prof. Juliette Lelieur (Section 3), and Prof. Milena Sterio (Section 4). All materials from the international colloquia have been published in the *Revue Internationale de Droit Pénal* (RIDP), in the below chronology:

- Section 1: Lorenzo Picotti & Beatrice Panattoni (Eds.), Traditional Criminal Law Categories and AI: Crisis or Palingenesis?, RIDP Vol. 94 issue 1, 2023;
- Section 3: Juliette Lelieur (Ed.), Artificial Intelligence and Administration of Criminal Justice, RIDP Vol. 94 issue 2, 2023;
- Section 2: Fernando Miró-Llinares, Constantin Duvac, Tudorel Toader & Mario Santisteban Galarza (Eds.), Criminalisation of AI-related offences;
- Section 4: forthcoming, as Vol. 96 issue 1, 2025.

PRÉFACE

Katalin Ligeti, John Vervaele et Gert Vermeulen[*]

Les systèmes de justice pénale du monde entier intègrent désormais bien souvent des technologies numériques basées sur l'intelligence artificielle (IA) dans la détection, l'enquête et la poursuite des infractions pénales. La promesse d'une analyse rapide et de résultats apparemment objectifs, basés sur une analyse scientifique, rend l'IA particulièrement attrayante pour le domaine juridique. La tentation est grande de croire que l'IA peut éliminer le "facteur humain", souvent perçu comme émotionnel, faillible et comme une menace potentielle à une prise de décision adaptée.

Pour exploiter ces avantages perçus, les développeurs de systems d'IA ont commencé à proposer leurs services aux autorités judiciaires et à d'autres acteurs de la chaîne pénale. Ils créent des algorithmes et des modèles statistiques exclusifs spécialement conçus pour la prévention, la répression et le jugement des infractions pénales.

Reconnaissant Au vu de l'impact profond de l'IA sur la justice pénale, l'AIPL a consacré son cycle scientifique 2020-2024 et son XXIe Congrès à l'examen l'étude des effets transformateurs de l'IA et des défis juridiques qu'elle pose pour dans le cadre dule droit pénal matériel et procéduralde la procédure pénale. Le sujet a été préparé abordé grâce par à un document de réflexion initiale et exploré par la suite lors de quatre colloques internationaux, chacun aboutissant à l'adoption d'une série de résolutions portant sur un aspect distinct du système de justice pénale.
Ce volume présente le document de réflexion de 2019 et les résolutions adoptées lors des colloques internationaux, structurées selon les quatre sections fondatrices de notre association :

- les résolutions portant sur les "Catégories traditionnelles de droit pénal et IA", adoptées lors du colloque de la Section 1 sur la partie générale du droit pénal, qui s'est tenu à Syracuse du 14 au 17 septembre 2022 ;
- les résolutions sur "Le droit pénal et la criminalisation face aux défis de l'IA", adoptées lors du colloque de la Section 2 sur la partie spéciale du droit pénal, qui s'est tenu à Bucarest du 14 au 16 juin 2023 ;

[*] Katalin Ligeti est présidente de l'AIDP/IAPL, doyenne de la faculté de droit, d'économie et de finance et professeure de droit pénal européen et international à l'Université du Luxembourg. John A.E. Vervaele est président honoraire de l'AIDP/IAPL, professeur émérite à l'Université d'Utrecht, Pays-Bas, et professeur de droit pénal européen au Collège d'Europe, Bruges, Belgique. Gert Vermeulen est directeur général des publications de l'AIDP/IAPL, rédacteur en chef de la RIDP et professeur titulaire de droit pénal européen et international, de droit pénal sexuel et de droit de la protection des données à l'Université de Ghent, Belgique.

- les résolutions sur "L'IA et l'administration de la justice pénale : Police prédictive, justice prédictive et preuves ", adoptées lors du colloque de la Section 3 sur la procédure pénale, qui s'est tenu à Buenos Aires du 28 au 31 mars 2023 ;
- les résolutions sur les "Perspectives internationales sur l'IA : défis pour la coopération judiciaire et le droit international humanitaire et pénal", adoptées lors du colloque de la Section 4 sur le droit pénal international, qui s'est tenu à Opatija les 7 et 8 décembre 2023.

Ces résultats scientifiques majeurs n'auraient pas été possibles sans le travail remarquable des quatre rapporteurs généraux : Lorenzo Picotti (Section 1), Fernando Miró (Section 2), Juliette Lelieur (Section 3) et Milena Sterio (Section 4). Tous les documents issus des colloques internationaux ont été publiés dans la *Revue Internationale de Droit Pénal* (RIDP) , selon la chronologie ci-dessous :

- Section 1 : Lorenzo Picotti & Beatrice Panattoni (Eds.), Traditional Criminal Law Categories and AI : Crisis or Palingenesis, RIDP Vol. 94 issue 1, 2023 ;
- Section 3 : Juliette Lelieur (Ed.), Intelligence artificielle et administration de la justice pénale, RIDP Vol. 94 issue 2, 2023 ;
- Section 2 : Fernando Miró-Llinares, Constantin Duvac, Tudorel Toader & Mario Santisteban Galarza (Eds.), Criminalisation of AI-related offences ;
- Section 4 : à venir, en tant que Vol. 96 numéro 1, 2025.

PREFACIO

Katalin Ligeti, John Vervaele y Gert Vermeulen[*]

Los sistemas de justicia penal, por todas partes del mundo, ya integran de forma rutinaria tecnologías informáticas basadas en la inteligencia artificial (IA) para la detección, la investigación y el enjuiciamiento de delitos. La promesa de la IA de realizar análisis rápidos y obtener resultados aparentemente objetivos y con base científica la hace especialmente atractiva en el ámbito jurídico. Existe una fuerte tentación de creer que la IA puede eliminar el "factor humano", a menudo percibido como emocional, falible y una amenaza potencial para la correcta toma de decisiones.

Para aprovechar estas ventajas percibidas, los desarrolladores de sistemas de IA han empezado a ofrecer servicios para las autoridades judiciales y otros actores de de la cadena penal. Crean algoritmos dedicados y modelos estadísticos diseñados específicamente para la prevención, la represión y el enjuiciamiento de delitos.

Dado el profundo impacto de la IA en la justicia penal, la AIDP dedicó su ciclo científico 2020-2024 y su XXIo Congreso a examinar los efectos transformadores de la IA y los retos jurídicos que plantea tanto para el derecho penal sustantivo como para el procesal. El tema se inició con un documento conceptual y se exploró a través de cuatro coloquios internacionales, cada uno de los cuales culminó con la adopción de un conjunto de resoluciones que abordaban un aspecto distinto del sistema de justicia penal.
Este volumen presenta el documento conceptual de 2019 y las resoluciones adoptadas , estructuradas según las cuatro secciones fundacionales de nuestra Asociación:

- las Resoluciones sobre "Categorías de Derecho Penal Tradicional y IA", adoptadas durante el coloquio de la Sección 1 sobre la parte general del Derecho Penal, celebrado en Siracusa del 14 al 17 de septiembre de 2022;
- las Resoluciones sobre "Derecho Penal y Criminalización ante los Desafíos de la IA", adoptadas durante el coloquio de la Sección 2 sobre la parte especial de Derecho Penal, celebrado en Bucarest del 14 al 16 de junio de 2023;
- las Resoluciones sobre "La IA y la Administración de Justicia Penal: Policía predictiva, justicia predictiva y pruebas", adoptadas durante el coloquio de la Sección 3 sobre Proceso Penal, celebrado en Buenos Aires del 28 al 31 de marzo de 2023;

[*] Katalin Ligeti es Presidenta de la AIDP/IAPL y Decana de la Facultad de Derecho, Economía y Finanza y Catedrática de Derecho Penal Europeo e Internacional de la Universidad de Luxemburgo. John A.E. Vervaele es Presidente Honorario de la AIDP/IAPL, Profesor Emérito de la Universidad de Utrecht, Países Bajos, y Profesor de Derecho Penal Europeo en el Colegio de Europa, Brujas, Bélgica. Gert Vermeulen es Director General de Publicaciones de la AIDP/IAPL, Redactor Jefe de la RIDP y Catedrático de Derecho Penal Europeo e Internacional, Derecho Penal Sexual y Derecho de Protección de Datos en la Universidad de Gante, Bélgica.

- las Resoluciones sobre "Perspectivas internacionales sobre la IA: desafíos para la cooperación judicial y el derecho internacional humanitario y penal", adoptadas durante el coloquio de la Sección 4 sobre Derecho Penal Internacional, celebrado en Opatija del 7 al 8 de diciembre de 2023.

Este importante resultado académico no habría sido posible sin la extraordinaria labor de los cuatro ponentes generales: Prof. Lorenzo Picotti (Sección 1), Prof. Fernando Miró (Sección 2), Prof. Juliette Lelieur (Sección 3) y Prof. Milena Sterio (Sección 4). Todos los materiales de los coloquios internacionales han sido publicado en en la *Revue Internationale de Droit Pénal* (RIDP), en la cronología que figura a continuación:

- Sección 1: Lorenzo Picotti & Beatrice Panattoni (Eds.), Traditional Criminal Law Categories and AI: Crisis or Palingenesis?, RIDP Vol. 94 issue 1, 2023;
- Sección 3: Juliette Lelieur (Ed.), Artificial Intelligence and Administration of Criminal Justice, RIDP Vol. 94 número 2, 2023;
- Sección 2: Fernando Miró-Llinares, Constantin Duvac, Tudorel Toader & Mario Santisteban Galarza (Eds.), Criminalización de los delitos relacionados con la IA;
- Sección 4: de próxima publicación, como Vol. 96 número 1, 2025.

CONCEPT PAPER
FOR THE 2020-2024 IAPL CYCLE

ARTIFICIAL INTELLIGENCE AND CRIMINAL JUSTICE

Katalin Ligeti[*]

Already in 1997, IBM supercomputer Deep Blue showed, to the astonishment of the world, that almost no activity was 'too human' for artificial intelligence (AI) systems, not even playing chess with and defeating the world chess champion Garry Kasparov. More than twenty years later, AI has become an integral part of our lives. From digital voice assistants like Siri and Alexa to automated purchase suggestions, from cleaning robots to drones, AI systems are everywhere and their diffusion is expected to grow exponentially in the future.

Unsurprisingly, AI-related projects and initiatives have mushroomed over the last few years. Just to name a few, in April 2018 the European Commission issued a communication titled 'Artificial Intelligence for Europe',[1] which was followed by the creation of a High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018. The UN Interregional Crime and Justice Research Institute (UNICRI) opened its Centre for Artificial Intelligence and Robotics in 2017, while the Committee of Ministers of the Council of Europe set up an Ad Hoc Committee on Artificial Intelligence in September 2019. In the context of the Council of Europe, it is also worth mentioning that the European Commission for the Efficiency of Justice (CEPEJ) came up with the 'European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment' at the end of 2018.[2]

AI raises a raft of questions for all legal systems around the world.[3] The first of these questions concerns the very same meaning of 'Artificial Intelligence' since there is no consensus on the exact meaning of this concept.[4] The definition by the European Commission can be used here as a useful reference point. According to the Commission, AI refers to

> systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-

---

[*] Katalin Ligeti is President of the AIDP/IAPL and Dean of the Faculty of Law, Economics and Finance and Professor of European and International Criminal Law at the University of Luxembourg. This concept paper has been conceived and drafted together with Dr. Fabio Giuffrida (University of Luxembourg). All webpages have been last accessed on 8 November 2019.

[1] COM(2018) 237 final, 25 April 2018. One year later, the Commission issued a new communication on 'Building Trust in Human-Centric Artificial Intelligence' COM(2019) 168 final, 8 April 2019.

[2] On AI and ethics see also, e.g., L. Floridi et al., 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 *Minds and Machines* 689.

[3] Among the uncountable studies see, for instance, W. Barfield and U. Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar 2018); A. Bensoussan and J. Bensoussan, *IA, robots et droit* (Bruylant 2019); T.F. Claypoole, *Law of Artificial Intelligence and Smart Machines: Understanding A.I. and the Legal Impact* (ABA Publishing 2019).

[4] For example, in S. J. Russel and P. Norvig, *Artificial intelligence: A modern approach* (3rd edn, Upper Saddle River: Pearson Education 2013) 1–5, eight definitions of AI are examined and compared.

> based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones […]).[5]

This definition singles out some intrinsic features of AI systems, namely their ability to: a) *collect* and *analyse data* from the surrounding environment; and b) *take actions* to achieve the machine's specific *goals*, which are usually predefined by a human operator. These abilities are the equivalent of what is usually meant with 'intelligence' (or 'rationality') when talking about human beings.[6] The reference to 'some degree of autonomy' is also of the utmost importance: despite human inputs, AI systems (can) act independently, e.g. they can choose among different courses of actions the one that looks the most appropriate to achieve their goals. Autonomy also refers to the fact that some AI systems, like humans, can *learn*: building both on the data they are fed with and on those they collect, AI systems can develop their 'skills' and adapt their 'behaviour' over time. 'Machine learning' thus implies that AI systems '[identify] patterns in available data and then [apply] the knowledge to new data'.[7]

Against this backdrop, legislators are required to keep pace with the scientific innovations and, when appropriate, regulate the unprecedented problems that AI raises. In the literature, there is already a considerable number of studies concerning the implications of AI for civil law, especially for liability law,[8] and criminal law literature is increasingly paying attention to the matter. There is indeed a strong need to conceptualise and address the several legal issues that AI poses. At the level of international organisations, while the EU has not yet launched any initiative concerning AI and criminal justice, the Council of Europe has recently established a Working Group of Experts on Artificial Intelligence and Criminal Law, which will mostly focus on substantive criminal law. On the basis of its work, the option of adopting a standard-setting instrument addressing AI, which might take the form of a Council of Europe convention, will be considered.[9]

By definition, criminal law rules deal mostly with human beings and their behaviours, so that the application of such rules to AI systems is not straightforward. For instance, to what extent is a crime committed by an AI system attributable to a human being? Perhaps stretching to sci-fi scenarios, what could be, if any, the conditions to consider AI systems themselves criminally responsible? These questions echo those concerning the criminal liability of legal persons, another sensitive topic with which legislators and courts have

---

[5] European Commission, 'Artificial Intelligence for Europe', cit., 1.

[6] Cfr. AI HLEG, 'A Definition of AI: Main Capabilities and Disciplines. Definition developed for the purpose of the AI HLEG's deliverables' (8 April 2019) available at https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence, 1).

[7] European Commission, 'Artificial Intelligence for Europe', cit., 11. 'Machine learning' is a broad category and includes different learning procedures: see more in, e.g., AI HLEG, 'A Definition of AI', cit., 3–4.

[8] See, for instance, A. Renda, 'Artificial Intelligence. Ethics, governance and policy challenges' (2019) Report of a CEPS Task Force, available at www.ceps.eu/ceps-publications/artificial-intelligence-ethics-governance-and-policy-challenges/, 82–90.

[9] www.coe.int/en/web/artificial-intelligence/work-in-progress.

had to cope in recent years and that was one of the subjects of the XX Congress.[10] Further issues to examine relate to the opportunities that AI systems present for the criminal justice system. Can we rely on AI to adopt decisions on criminal law cases? How can AI systems help law enforcement authorities in preventing, detecting, and combating crime?

The XXI International Congress of Penal Law will try to answer these and further questions stemming from the interplay between AI and criminal law by taking into account the various aspects of criminal justice: general and special part of substantive criminal law, procedural law as well as issues linked to the administration of justice, and international criminal law. The four sections of the Congress should analyse changes and tendencies regarding policies, norms, and practices. The technical nature of the subject and the indisputable fact that 'the technological developments have far outpaced legal or policy debates'[11] around it call for an *inter-disciplinary* approach. Practitioners, scholars from other branches of law, and experts in fields other than law, especially those involved in developing AI systems, but also bioethicists, criminologists, scientists, and crime analysts, should be encouraged to attend the Congress and share their expertise. Cross-fertilisation of ideas is crucial to understanding the multifarious aspects of AI and pave the way for mature reflections on how (criminal) law should deal with such a complex matter.

Section 1. Traditional Criminal Law Categories and AI: Crisis or Palingenesis?

Section 1 of the Congress will focus on the general part of substantive criminal law and address the question of whether and how traditional criminal law categories – especially *actus reus*, *mens rea*, and causation – can apply to crimes committed by/through AI systems. When AI crosses the path of criminal law, these traditional concepts may experience a crisis. The example of autonomous driving is helpful to grasp the reasons of this.[12] In the event of an accident involving personal injury or causing death to a passer-by, who is responsible? Several options can be explored. The most unrealistic, at least for the time being, is that of considering the car itself (criminally) responsible. Although some authors do not entirely rule out the possibility of endorsing a direct liability model,[13] this

[10] When assessing the impact of AI on substantive criminal law, parallels are often drawn between corporate criminal liability and AI's (potential) criminal liability (see, e.g., U. Pagallo and S. Quattrocolo, 'The impact of AI on criminal law, and its twofold procedures', in W. Barfield and U. Pagallo (eds.), op. cit., 402–405).

[11] A. G. Ferguson, 'Policing Predictive Policing' (2017) 94 *Washington University Law Review* 1115, 1148.

[12] The issue of autonomous driving is also the main focus of the Council of Europe Working Group of Experts on AI and Criminal Law. On criminal law issues related to autonomous driving see, e.g., J. Gurney, 'Driving into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles' (2015) 5 *Wake Forest Journal of Law & Policy* 393; S. Gless et al., 'If Robots Cause Harm, Who Is to Blame: Self-Driving Cars and Criminal Liability' (2016) 19 *New Criminal Law Review* 412. For broader remarks beyond autonomous driving see, e.g., F. Basile, 'Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine' (2019) *Diritto Penale e Uomo*, available at https://dirittopenaleuomo.org/contributi_dpu/intelligenza-artificiale-e-diritto-penale-quattro-possibili-percorsi-di-indagine/, 1, 24 ff.

[13] See G. Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* (Springer 2015) 102 ff.

seems unfeasible. Even admitting that the accident caused by the car amounts to an *actus reus*, it would be very difficult to claim that this act was supported by the car's *mens rea*:

> It would make little (social) sense to attribute culpability to a being that is incapable of recognizing its own past and evaluating its past actions in accordance with a moral reference system. An entity that does not have a conscience cannot participate in a dialogue on ethical issues and cannot respond to reproach.[14]

Likewise, it is even less sensible to 'punish' a machine, at least as long as the machine 'is not imbued with a will to live'.[15] In other words, if the AI system is not in a position of understanding the sanction and learning from it, punishment is of no use.[16]

In order not to create accountability gaps, it would then be necessary to look for the human responsibility behind the accident. This would imply ascertaining whether the manufacturer, the programmer, and/or the user are responsible. The easiest scenario would be that of an autonomous car that is expressly programmed with the aim of killing, as in this case the AI system would simply be used as an instrument of crime.[17] Leaving aside this somehow extreme hypothesis, however, some problems arise. The assessment of the human responsibility should in fact take into account several factors, e.g. whether there was any negligence in designing/using the car and whether a human being (in the car or remotely) was able to intervene and disengage the autonomous driving system.[18] For example, the US National Transportation Safety Board has recently found that, in the accident that was caused in 2018 by an Uber self-driving test vehicle in Arizona and that killed a woman who was crossing the road, there were some flaws in the system, which was not in a condition to recognise a person walking outside pedestrian crossing. Furthermore, the driver insider the car was distracted when the accident happened and she could thus face criminal charges.[19]

---

[14] S. Gless et al., 'If Robots Cause Harm', op. cit., 423.

[15] *Ivi*, 424, where the authors underline that it is difficult to imagine sanctions 'against Intelligent Agents that would fulfill the same purposes as criminal sanctions imposed on human beings', since robots 'are incapable of understanding the meaning of punishment and therefore cannot draw a connection between anything "done to them" and their prior fault'.

[16] For similar remarks on the little sense of 'punishing' AI machines – at least for the time being – see, for instance, U. Pagallo, *The Laws of Robots. Crimes, Contracts, and Torts*, Springer, 2013, 50–51; D. Lima, 'Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law' (2018) 69 *South Carolina Law Review* 677, 688–689; T. C. King et al., 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions' (2019) *Science and Engineering Ethics* 1, 20.

[17] See, S. Gless et al., 'If Robots Cause Harm', op. cit., 425; G. Hallevy, 'The Basic Models of Criminal Liability of AI Systems and Outer Circles' (2019) available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3402527, 2–3.

[18] See, more extensively, F. Douma and S.A. Palodichuk, 'Criminal Liability Issues Created by Autonomous Vehicles' (2012) 59 *Santa Clara Law Review* 1157, 1160 ff. For further remarks on negligence and liability for crimes committed by AI systems see S. Beck, 'Intelligent agents and criminal law—Negligence, diffusion of liability and electronic personhood' (2016) 86 *Robotics and Autonomous Systems* 138.

[19] 'Uber in fatal crash had safety flaws say US investigators' (*BBC News*, 6 November 2019) available at www.bbc.com/news/business-50312340.

The liability models that can be used to attribute the responsibility for AI machines' accidents to human beings are the 'perpetration-by-another' model, whereby the AI system is considered the 'other' entity that humans use to commit the crime, and the 'natural probable consequence' model, according to which the manufacturer, programmer, and/or user are responsible because the offence is a natural and probable consequence of their (negligent) action of creating, programming, and/or using the machine.[20] These categories are to be found in several criminal justice systems, for instance to regulate the criminal responsibility of accomplices, but their applicability to AI systems, which are different from both mere instruments of crime and (human) partners in crime, deserves further reflections. By the same token, it is to be examined whether strict liability models can play a role in this context and whether an agreement can be found on the notion of a socially permissible risk concerning autonomous driving, since 'the crucial question in the development of automated driving might concern what kind of risk respective societies are willing to accept'.[21]

In addition, another traditional concept of criminal law, i.e. causation, may have to be rethought when it comes to AI-related crimes. It can happen that offence committed by AI machines cannot be easily traced back to the human being behind the system. For instance, one could think of robots that were produced by humans who had no intent whatsoever to commit a crime. If the accident is caused by the faulty process of machine learning that the AI system undertakes, rather than by a potential human negligence in programming or using it, should we consider the causation chain between the human behaviour and the accident to be interrupted by an unpredictable event? Or should not we think in this way since AI systems cannot be considered as proper 'persons' who can break the chain of causation?[22]

In sum, the attribution of crimes committed by/through AI systems to responsible individuals is a major challenge to traditional ways of criminal law thinking. This section of the Congress should thus examine the consequences of AI for the well-established categories of the general part of criminal law, especially *mens rea*, *actus reus*, and causation, and discuss whether they are sufficient to regulate the new phenomena or need instead some (deep) rethinking to face the challenges ahead.

Section 2. Old and New Criminal Offences: AI Systems as Instruments and Victims

Section 2 of the Congress will focus on the special part of substantive criminal law, which is likely to undergo substantial changes in the coming years due to the advent and diffusion of AI. Section 2 will examine at least two different scenarios. First, it should discuss

[20] Cfr. D. Lima, op. cit., 691 ff.; G. Hallevy, 'The Basic Models of Criminal Liability', op. cit., 1–8; T. C. King, op. cit., 20–22; P. Yeoh, 'Artificial intelligence: accelerator or panacea for financial crime?' (2019) 26 *Journal of Financial Crime* 634, 638–640.

[21] S. Gless, 'Working Paper II. Document prepared for the 1st meeting of the Working Group of Experts on Artificial Intelligence and Criminal Law' (2019) available at www.coe.int/en/web/cdpc/home, 4.

[22] Cfr U. Pagallo, op. cit., 53 and 75. See also D. Lima, op. cit., 684.

how AI systems can be used to commit 'traditional' crimes. Some studies have already highlighted the extent to which criminal organisations can benefit from AI. For instance, drug trafficking may become easier – and much less risky for criminals – if the illegal substances are moved from one place to another by means of drones.[23] The same goes for terrorist attacks that may be carried out by placing explosive materials on AI machines.[24] Another crime that AI may facilitate is online fraud, especially 'spear phishing', which refers to 'email or electronic communications scam targeted towards a specific individual, organization or business'.[25] While phishing by means of emails that are blatantly fake is not often successful, AI systems can create and send fraudulent emails that are tailored to the recipient, who can then be convinced to follow a malicious link and/or share his or her data with the fraudster.[26]

Second, AI may lead to *new* crimes altogether. In a 2019 report by UNICRI and Interpol, we read that a 'study on "new crimes" involving the malicious use of AI and robotics should be conducted'.[27] On the one hand, AI systems can become 'victims' of crime and it is likely that new definitions and rules will be needed to regulate these situations.[28] For instance, AI systems may be sabotaged by third parties so that these systems will be impaired from achieving their goals and/or induced to commit a crime. One could think of persons who intentionally disrupt the software of autonomous driving cars, in this way provoking accidents that were entirely out of control of the programmer and user of the vehicle.[29]

On the other hand, AI is a powerful instrument for dangerous behaviours that could be criminalised in the future. AI systems may be tasked, for example, with the creation and spreading of fake news.[30] While this already represents a complex issue in contemporary

[23] According to Europol, organised crime groups 'involved in drug trafficking will likely invest in drone technology for trafficking purposes in order to avoid checks at border crossing points, ports and airports' (Europol, 'European Union Serious and Organised Crime Threat Assessment' (2017) 34).

[24] See F. Douma and S.A. Palodichuk, op. cit., 1166; M. Brundage et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (2018) Future of Humanity Institute, University of Oxford; Centre for the Study of Existential Risk, University of Cambridge, available at https://maliciousaireport.com/, 27 ff.; T. C. King et al., op. cit., 12–13.

[25] This definition, which is taken from the website of the cybersecurity and anti-virus provider Kaspersky (www.kaspersky.com/resource-center/definitions/spear-phishing), also adds: 'Although often intended to steal data for malicious purposes, cybercriminals may also intend to install malware on a targeted user's computer'.

[26] Spear phishing has been subject to an experiment by two computational social scientists, J. Seymour and P. Tully, 'Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter' (2016) available at www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf, which is discussed by T. C. King et al., op. cit., 2. Risks of spear phishing are also examined by, e.g., M. Brundage et al., op. cit., 18–21.

[27] UNICRI Centre for Artificial Intelligence and Robotics and Interpol Innovation Centre, 'Artificial Intelligence and Robotics for Law Enforcement' (2019) available at www.unicri.it/in_focus/on/interpol_unicri_report_ai, 23.

[28] Cfr. F. Basile, op. cit., 32–33.

[29] F. Douma and S.A. Palodichuk, op. cit., 1165; M. Brundage et al., op. cit., 5.

[30] M. Brundage et al., op. cit., 29 and 46.

society, similar conducts do not usually amount to a crime, with a few exceptions.[31] Since AI has the potential to escalate this phenomenon to the point where it would represent a daunting and unprecedented threat to our democracies, as the machines' level of accuracy is likely to make it difficult even for the most attentive user to distinguish truth from fiction, it is worth examining whether the conducts at issue should be criminalised in order to reduce their potentially devastating effects.

Finally, the possible interactions between AI and cryptography, with a focus on those technologies that build on cryptography such as blockchain, will deserve further attention in the future. It is difficult to regulate the legal implications – including criminal law ones – of blockchain and crypto-assets per se. The possible combination with AI may raise even more complex questions as this may facilitate the commission of existing or new crimes, and potentially require the introduction of *ad hoc* criminal law provisions. Due to the lack of any in-depth analysis of the issue in the literature, the Congress will represent the ideal opportunity to start identifying these forthcoming challenges and reflecting on them.

Section 3. AI and Administration of Justice: Predictive Policing and Predictive Justice

Section 3 will examine the impact of AI on the administration of justice. In particular, it will focus on criminal procedural law and, more broadly, law enforcement, by looking at predictive policing and predictive justice mechanisms. By using algorithms that process enormous quantity of data, these mechanisms make predictions about where and when crimes are likely to be committed, and even by whom in some cases (predictive policing)[32] and about whether a suspect or defendant is likely to flee or commit further crimes, with the consequence that criminal courts can deny bail or opt for harsh sentences (predictive justice). These are far from being sci-fi speculations: already in 2006, a US scholar argued that 'prediction of criminality has become de rigueur in our highly administrative law enforcement and prison sectors—seen as a necessity, no longer a mere convenience'.[33] More recently, during the 2018 Global Meeting on the Opportunities and Risks of AI and Robotics for Law Enforcement, 'the use of AI tools for the purposes of prediction and analysis'[34] turned out to be the most cited application of AI technology for law enforcement purposes.

---

[31] See, e.g., A. Schetzer, 'Governments are making fake news a crime – but it could stifle free speech' (*The Conversation*, 7 July 2019) available at https://theconversation.com/governments-are-making-fake-news-a-crime-but-it-could-stifle-free-speech-117654.

[32] Traditionally, predictive policing 'is not actually predicting a particular crime, but predicting an elevated risk of crime based on pre-determined place-based factors', but there is now a shift towards 'the use of predictive technologies to identify individuals and groups involved in predicted criminal activity' (A. G. Ferguson, 'Policing Predictive Policing', op. cit., 1142).

[33] B.E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age* (University of Chicago Press 2006) 16.

[34] UNICRI Centre for Artificial Intelligence and Robotics and Interpol Innovation Centre, op. cit., 9. See also C. Slobogin, 'Principles of Risk Assessment: Sentencing and Policing' (2018) 15 *Ohio State Journal of*

In the US, for instance, Californian police use a software called PredPol to '[p]redict where and when specific crimes are most likely to occur',[35] although this instrument has recently been met with criticism as it did not help in reducing crime.[36] Some European police forces resort to similar software, Precobs (Pre Crime Observation System).[37] The logic behind these and other predictive policing systems is simple: some crimes, such as theft and robberies, 'are to a large extent predictable, because criminals with a distinguishable profile tend to commit the same type of crime, at roughly the same location and time of the day'.[38]

As for predictive justice, there are nowadays reportedly 'more than 200 risk assessment tools available in criminal justice and forensic psychiatry, which are widely used to inform sentencing, parole decisions, and post-release monitoring'.[39] One of the most famous, at least in the US, is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions).[40] COMPAS assesses the risk of recidivism, which is calculated by taking into account both an interview with the defendant and information from his or her criminal history. The COMPAS risk assessment, however, 'does not predict the specific likelihood that an individual offender will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a *similar data group*'.[41]

These words are taken from a landmark decision in the field of AI and criminal justice, *State v. Loomis* (2016), where the use of COMPAS was challenged before the Supreme Court of Wisconsin. The defendant, who was sentenced to six years of imprisonment after the COMPAS risk assessment had considered his risk of recidivism high, claimed that his right to due process had been violated because, inter alia, it was unclear how COMPAS made its assessments, and it was therefore impossible to challenge their accu-

Criminal Law 583; M. Gialuz, 'Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre di *risk assessment tools* tra Stati Uniti ed Europa' (2019) *Diritto penale contemporaneo* 1, available at www.penalecontemporaneo.it/d/6702-quando-la-giustizia-penale-incontra-l-intelligenza-artificiale-luci-e-ombre-dei-risk-assessment-too.

[35] www.predpol.com/.

[36] M. Puente and C. Chang, 'LAPD changing controversial program that uses data to predict where crimes will occur' (*Los Angeles Time*, 15 October 2019) available at www.latimes.com/california/story/2019-10-15/lapd-predictive-policing-changes.

[37] See A. Završnik, 'Algorithmic justice: Algorithms and big data in criminal justice settings' (2019) *European Journal of Criminology* 1, 2.

[38] R. Peeters and M. Schuilenburg, 'Machine justice: Governing security through the bureaucracy of algorithms' (2018) 23 *Information Polity* 267, 272.

[39] *Ivi*, 273.

[40] See, for instance, A. Christin, 'Algorithms in practice: Comparing web journalism and criminal justice' (2017) *Big Data & Society* 1, 5–6.

[41] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), para. 15 (emphasis added). Harcourt speaks of predictive policing instruments as 'actuarial methods', as 'they use statistical methods […] on large datasets of criminal offending rates in order to determine the different levels of offending associated with a group or with one or more group traits and, on the basis of those correlations, to predict the past, present, or future criminal behavior of a particular person' (B. E. Harcourt, op. cit., 16).

racy, and the use of the predictive justice software had violated his right to an individualised sentence.[42] The Supreme Court of Wisconsin did not share his views and decided that, if used properly, courts' reliance on COMPAS risk assessments in the sentencing phase does not violate the right to due process.[43] In the case of Loomis – and this should happen, according to the Supreme Court of Wisconsin, in any other case where predictive justice instruments are used – the court of lower instance reached its decision by relying on 'other independent factors', so that the use of the COMPAS risk assessment was 'not determinative in deciding whether Loomis could be supervised safely and effectively in the community'.[44]

In the light of the foregoing, section 3 of the Congress shall delve into the several problems that predictive policing and justice instruments raise for the administration of justice. First, it should be discussed whether predictive policing is not in fact counterproductive. As predictions on the future are made on the basis of data from the past, the algorithms can lead police authorities to invest their money and resources in patrolling areas that are already known to be prone to crime, while all other areas and crimes (including those offences for which the reporting rate is low) could continue to be neglected.[45] One of the reasons why PredPol attracted criticism was precisely because it 'essentially provided information already being gathered by officers patrolling the streets'.[46]

Second, if a person is suspected of committing future crimes – and then investigated – on the basis of algorithmic calculations that draw on statistical data and/or the analysis of patterns and behaviours that are not criminal per se, some basic human rights would be at stake, beginning with the presumption of innocence. Incidentally, this might also exacerbate the relations between the public and law enforcement authorities, which, especially in some areas, are already tense and rife with mistrust.[47] Third, predictive policing and justice are thought to provide neutral and objective information, while human judgements are intrinsically biased. This argument has been rebutted by studies that proved that AI machines used in the administration of justice 'embed existing biases and

[42] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), para. 34. For a commentary see H.-W. Liu, C.-F. Lin and Y.-J. Chen, 'Beyond *State v Loomis*: artificial intelligence, government algorithmization and accountability' (2019) 27 *International Journal of Law and Information Technology* 122.

[43] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), para. 8.

[44] *Ivi*, para. 9. The Court warned about some risks connected with the use of COMPAS, yet this is unlikely to be sufficient to eradicate all the problems stemming from risk assessments (see 'Criminal Law – Sentencing Guidelines – Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing—*State v. Loomis*, 881 N.W.2d 749 (Wis. 2016)' (2017) 130 *Harvard Law Review* 1530, 1536).

[45] R. Peeters and M. Schuilenburg, op. cit., 274; A. Završnik, op. cit., 7.

[46] M. Puente and C. Chang, 'LAPD changing controversial program', op. cit.

[47] A. G. Ferguson, 'Policing Predictive Policing', op. cit., 1163.

perpetuate discrimination'.[48] After all, since AI systems work on the basis of data inputted by human beings, the choice of these data becomes crucial and may turn out to be itself biased.[49] The human component can never be entirely set aside also because algorithms usually come up with a number or a given result, but it is then for the user to attach a meaning to that figure or outcome: 'For instance, at what probability of recidivism should a prisoner be granted parole? Whether this threshold ought to be a 40 percent or an 80 percent risk of recidivism is an inherently "political" decision based on the social, cultural and economic conditions of the given society'.[50]

Finally, predictive policing and justice prompt broader systematic reflections on the future role of public authorities (courts, prosecutors, and police) in the enforcement of the (criminal) law, a role that will become much more proactive compared to the (mostly) reactive one they currently play.[51] Furthermore, as their activities are likely to be always more influenced, if not determined, by mathematical formulas,[52] we could witness a silent shift of responsibility from public authorities towards (private) companies, and ultimately towards the experts who create and programme AI systems. This is however highly problematic from the perspective of public authorities' accountability and transparency,[53] especially because the way AI systems work is often not clear at all, and it may also be covered by trade secret.[54] Many sensitive decisions concerning individuals are thus left in the hands of obscure and unclear mechanisms ('black-box AI').[55] The scenario can become even more problematic if AI will be used not only to predict future crimes or risks of recidivism but also to decide criminal law cases altogether, replacing judges and juries. Automated decision systems have already been tested in civil proceedings so

---

[48] A. Završnik, op. cit., 4, who refers to the ProPublica's report by J. Angwin et al., 'Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks' (2016) available at www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. See also R. Peeters and M. Schuilenburg, op. cit., 274; Council of Europe Committee of experts on internet intermediaries (MSI-NET), 'Study on the human rights dimension of automated data processing techniques (in particular algorithms) and possible regulatory implications' (2017) MSI-NET(2016)06 rev3 FINAL, 11–12.

[49] A. Završnik, op. cit., 8–9. See also A. Christin, op. cit., 3. Humans can also commit errors and this may affect the quality of the predictive mechanism (see A. G. Ferguson, 'Policing Predictive Policing', op. cit., 1150 ff.).

[50] A. Završnik, op. cit., 10.

[51] See, for instance, A. G. Ferguson, 'Predictive Prosecution' (2016) 51 *Wake Forest Law Review* 705, 731 ff.

[52] Cfr. R. Peeters and M. Schuilenburg, op. cit., 274–275.

[53] See A. G. Ferguson, 'Policing Predictive Policing', op. cit., 1169 ff.; A. Babuta, M. Oswald and C. Rinik 'Machine Learning Algorithms and Police Decision-Making. Legal, Ethical and Regulatory Challenges' (2018) RUSI Whitehall Report 3-18, 17–22.

[54] See, for instance, P. W. Nutter, 'Machine Learning Evidence: Admissibility and Weight' (2019) 21 *Journal of Constitutional Law* 919, 941–944.

[55] See the 'Statement of Concern About Predictive Policing by ACLU and 16 Civil Rights Privacy, Racial Justice, and Technology Organizations' (2016) published on *American Civil and Liberties Union* (https://www.aclu.org); A. Christin, op. cit., 3. As the EU High-Level Expert Group on AI explains, 'Some machine learning techniques, although very successful from the accuracy point of view, are very opaque in terms of understanding how they make decisions. The notion of *black-box AI* refers to such scenarios, where it is not possible to trace back to the reason for certain decisions' (AI HLEG, 'A Definition of AI', cit., 5).

it would not be surprising if there will be some attempts to inquire whether they can also work in criminal law contexts.[56]

In sum, extensive and in-depth reflections on whether, to what extent and under which conditions predictive methods are truly compatible with the basic tenets of modern democracies – including fundamental rights such as privacy, presumption of innocence, and defence rights – cannot be postponed anymore.

Section 4. International Perspectives on AI: Challenges for Judicial Cooperation and International Humanitarian/Criminal Law

Section 4 of the Congress will examine some international implications of the use of AI. In particular, this section will deal with the impact of AI on: a) evidence gathering, which will be looked at through the prism of international cooperation; and b) international humanitarian law and international criminal law, especially with regard to the use of robots in war contexts.

As for evidence gathering, it ought to be noted that AI systems can be of great value to law enforcement authorities, even beyond the above-mentioned examples of predictive policing.[57] Analysing, for example, DNA or social media profiles 'produces large amounts of complex data in electronic format',[58] which may contain useful patterns that human analysis could not be able to grasp. AI-backed tools can also be used to identify fake art works[59] or persons by means of facial recognition software, which 'could identify a defendant even with video or photographic evidence in less than ideal circumstances'.[60] AI can also help in locating events and places. In 2017, the International Criminal Court requested the arrest of a Libyan warlord by relying on information deriving from satellite images and videos, which were uploaded online by his acolytes and showed some executions he had ordered: 'Geographical features seen in the videos—buildings, roads, trees, hills—were located via time-stamped high-resolution satellite images. In this way,

[56] See F. Basile, op. cit., 14–16.

[57] See, for instance, L. Goldmeier, 'How Artificial Intelligence is Revolutionizing Investigation for Law Enforcement' (*Briefcam*, 21 August 2018) available at www.briefcam.com/resources/blog/how-artificial-intelligence-is-revolutionizing-investigation-for-law-enforcement/. An extensive overview of the ways AI can support law enforcement can be found in the recent report by UNICRI Centre for Artificial Intelligence and Robotics and Interpol Innovation Centre, op. cit. A testament to the increasing importance of the topic is the fact that the 2019 OSCE Annual Police Experts Meeting was devoted to 'Artificial Intelligence and Law Enforcement - An Ally or Adversary?' (see www.osce.org/event/2019-annual-police-experts-meeting).

[58] C. Rigano, 'Using Artificial Intelligence to Address Criminal Justice Needs' (2019) *National Institute of Justice Journal*, 6.

[59] L. Floridi, 'Artificial Intelligence, Deepfakes and a Future of Ectypes' (2019) 31 *Philosophy & Technology* 317. Interestingly, the author notes that AI can also be used to *create* fake work arts.

[60] P. W. Nutter, op. cit., 929–930.

video, photos, satellite images, and other data are triangulated to verify events in a specific time and place'.[61] While in that case most of the analysis was carried out by humans, in the future 'substantial portions of it could be automated or enhanced by machine learning'.[62]

When the outcome of algorithmic calculations by AI systems is used as evidence before a criminal court, however, the fundamental right to a fair trial risks being violated at least for two different reasons. First, as mentioned, the algorithmic processes that analyse the data and end up providing public authorities with a given piece of evidence are often obscure, so that the defendant is not in a position to challenge the way in which evidence has been gathered: 'Insofar as individuals in a legal process are unable to understand and contest, even with the help of legal counsel, complex algorithmic systems used to process evidence alleged to relate to them, there is a significant threat to due process rights'.[63]

Second, and consequently, the use of AI-related evidence poses a risk to the principle of equality of arms.[64] Even if this principle has to discount the difference between the situation of public authorities and that of individuals, an insurmountable advantage to the former flows from the use of AI in the process of evidence gathering. If investigations are based on AI techniques, therefore, the defendant should be in a position to understand how evidence has been gathered, while 'the denial of discovery in relation to the program, code, or data governing the AI system […] would represent a clear infringement of the principle of Equality of Arms between the parties'.[65] If the code is discovered, the defendant will likely need to find an expert who would be able to understand and challenge the algorithmic process on which police and prosecutors relied.[66] At the same time, however, the integral discovery of how AI machines work may be detrimental to law enforcement authorities' activities and companies' trade secrets, and may also lead to endless disputes over the reliability of AI systems that could hamper or substantially prolong criminal proceedings.

In sum, it will be necessary to strike a balance between the advantages that AI brings to the administration of justice and the respect of key principles of criminal justice, such as the right to due process and rights of defence, which are at stake when an individual is left to argue against obscure decisions that are in essence taken by AI experts outside the

---

[61] S. Livingston and M. Risse, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' (2019) 33 *Ethics & International Affairs* 141, 143.

[62] M. M. Maas, 'International law does not compute: Artificial intelligence and the development, displacement or destruction of the global legal order' (2019) 20 *Melbourne Journal of International Law* 29, 44.

[63] M. Veale, 'Algorithms in the Criminal Justice System' (2019) The Law Society of England and Wales, 57. Cfr. S. Gless, 'Working Paper II', op. cit., 4–5.

[64] *Ibidem.*

[65] U. Pagallo and S. Quattrocolo, op. cit., 396.

[66] In this case, the traditional criteria to evaluate scientific evidence – such as the known US *Daubert* criteria – can come into play in order to assess whether the algorithm possesses some sufficient level of accuracy (P. W. Nutter, op. cit., 948).

walls of criminal courts.[67] Since AI continues to lag in common sense reasoning, thereby profoundly questioning the tenets of criminal procedure, session 4 of the Congress should examine whether AI-backed tools should be hold to a certain standard of explanation and if yes what is the applicable standard and what are the guarantees that should surround the use of AI-related evidence. For instance, a solution that could help foster reliability and transparency of AI techniques would be, according to some authors, 'to ask (and provide) for independent certification of the AI system's trustworthiness. An expert-witness could be appointed by the judge to verify either the algorithmic process, or the neural network of a certain AI system, whenever the parties express their doubt about the correctness of automated data'.[68]

While new approaches and solutions to evidentiary matters are needed at the national level, the situation becomes even more complex in cross-border settings. Cross-border exchange of evidence, especially from the perspective of the admissibility and use of evidence in a different State than that in which evidence was gathered, has always represented a critical issue of international cooperation in criminal matters. Even in a context such as that of the European Union, where harmonisation in criminal matters is on the rise, there has been so far no political will to agree on minimum rules concerning the mutual admissibility of evidence.[69] On top of that, the new – and still largely unresolved – problems connected with digital evidence add a further note of complexity.[70] Against this backdrop, therefore, it is an open question, which has not been yet addressed in the literature, whether the existing instruments of cooperation in criminal matters can ensure exchange, admissibility, and use of AI-related evidence in a satisfactory way.[71] If each country ends up regulating the issue of AI and criminal evidence according to its own principles, rules, and perhaps even technical standards, the panoply of different regimes

[67] See P. W. Nutter, op. cit., *passim*; A. Završnik, op. cit., 14.

[68] U. Pagallo and S. Quattrocolo, op. cit., 398. The authors however notice that, while this 'would certainly increase the chances to challenge the accuracy of the data', it only represents 'an "indirect" challenge, since it would be mediated by the direct experience of the court's expert, whom the defence may not trust' (*ibidem*). See also M. Cross, '*Algorithms and Schrodinger's Justice*' (2017) *The Law Society Gazette*. In this context, it is worth adding that Principle 4 of the CEPEJ Ethical Charter is that of 'transparency, impartiality and fairness', according to which data processing methods should be made accessible and understandable, and external audits should be authorised.

[69] See, for instance, J. Vervaele, 'Lawful and Fair Use of Evidence from a Human Rights Perspective', in F. Giuffrida and K. Ligeti (eds.), *Admissibility of OLAF Final Reports as Evidence in Criminal Proceedings* (University of Luxembourg 2019) 56–67.

[70] Once more, the intense negotiations on 'e-evidence' within the EU are a testament to these new challenges. See, for instance, S. Tosza, 'The European Commission's Proposal on Cross-Border Access to E-Evidence. Overview and Critical Remarks' (2019) *eucrim* 212. The topic was also addressed during the XIX International Congress of Penal Law (see section IV of the Recommendations of that Congress).

[71] See also S. Gless, 'Working Paper II', op. cit., 5–6, where the author points out that the Council of Europe Convention on Cybercrime may not be sufficient to face all the challenges connected with AI-related evidence.

may hamper judicial cooperation, so that one may wonder whether a coordinated approach on the international level would not be appropriate.[72]

As in any other case where AI systems may be used, however, the positive effects of the new technologies should not be forgotten.[73] It is worth mentioning that, while it brings international cooperation in uncharted territory, AI could also help national authorities to deal more efficiently with requests for cooperation. According to UNICRI and Interpol, one example of possible future use of AI and robotics consists precisely in autonomously researching, analysing and responding to requests for international mutual legal assistance.[74]

Moving on to international humanitarian law (IHL) and international criminal law (ICL), the issue of autonomous weapon systems (AWSs) and their impact on traditional principles of IHL and ICL has gained attention in the literature. Governments invest massively in research and realisation of AWSs, which, once fully created and extensively diffused, can represent invaluable resources for the military. An AWS can be defined as 'a weapon system that, based on conclusions derived from gathered information and preprogrammed constraints, is capable of *independently selecting and engaging target*'.[75] 'Autonomous' is therefore different from 'automated', since only 'autonomous weapons' can act independently of human inputs. A difference is usually made between 'human-*out*-of-the-loop' weapons, which are indeed the 'autonomous' ones, and 'human-*in*-the-loop' or 'human-*on*-the-loop' weapons, which instead feature some form of human control.[76] For the purpose of this paper, 'AWSs', 'killer robots', and 'AI systems' will be used as synonyms.

The first question AWSs raise is not strictly legal but has noteworthy legal implications: can their use make war a 'less serious issue' and therefore cause more wars than in the past? It is unquestionable that the use of robots by a given State reduces the number of its own losses in war.[77] A robot-war can 'lower public awareness' since 'a fully-automated

---

[72] Further problems might also be related to the issue of dual criminality, a traditional principle of mutual legal assistance, e.g. if some countries allow the use of self-driving cars and others do not (*ibidem*).

[73] For instance, S. Gless et al., 'If Robots Cause Harm', op. cit., 430–431, stress that, in spite of the complex problems that autonomous driving raises, society may nonetheless have 'a valid interest in promoting the use of self-driving cars' as they 'might indeed reduce the overall harm caused in street traffic'.

[74] UNICRI Centre for Artificial Intelligence and Robotics and Interpol Innovation Centre, op. cit., vi.

[75] R. Crootof, 'The Killer Robots Are Here: Legal and Policy Implications' (2015) 36 *Cardozo Law Review* 1835, 1854 (emphasis added).

[76] See, e.g., Human Rights Watch, 'Losing Humanity. The Case against Killer Robots' (2012) available at www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots, where the following three definition can be found: 'Human-*in*-the-Loop Weapons: Robots that can select targets and deliver force only with a human command; Human-*on*-the-Loop Weapons: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots' actions; and Human-*out-of*-the-Loop Weapons: Robots that are capable of selecting targets and delivering force without any human input or interaction'. See also, for instance, P. Alston, 'Lethal Robotic Technologies: The Implications for Human Rights and International Humanitarian Law' (2011) 21 *Journal of Law, Information & Science* 35, 40–41.

[77] M. Wagner, 'The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems' (2014) 47 *Vanderbilt Journal of Transnational Law* 1, 10.

military mission transforms war into a fairly *technical and bureaucratic* operation, risk-free so to speak, so that causes of war may also be trivial, once you imagine both armies engaging no humans but only robot soldiers'.[78] To put it even more bluntly, 'a president who sends someone's son or daughter into battle has to justify it publicly … But if no one has children in danger, is it a war?'.[79] In essence, AWSs can change the approach of politicians and public opinion to war, in a way that does not necessarily help to reduce wars in the future, rather the contrary. The impact of AWSs on the *ius ad bellum*, namely the set of rules that regulate the conditions to enter into war, deserves therefore further attention.[80]

Second, AWSs can also affect the *ius in bello*, which instead refers to the principles and rules that should apply during war.[81] There is an ongoing debate on whether AWSs undermine or strengthen the fundamental principle of distinction, according to which no civilians or civilian targets can be attacked during wars. On the one hand, one may argue that, as long as AWSs are programmed to avoid civilian targets, they may actually be better placed than human combatants to ensure the respect of the principle at hand.[82] On the other, however, there is no guarantee that, in practice, robots will be able to spare more civilians than human beings can, not least because AI systems – even the most advanced ones – will not have the necessary human abilities to figure out whether, in a given situation, a person or a target is civilian or not.[83] The following example is enlightening:

> During a counterinsurgency operation in a village, soldiers receive information that combatants may be hiding inside a house. Unbeknownst to the soldiers, no insurgents are present. Inside of the home, boys are playing with a ball. The children kick the ball towards the gate as the soldiers enter the main door. The male inhabitants of this area carry a dagger called the *kirpan* for purely religious reasons. One of the parents watching the children realizes that the children are in danger and tries to warn them by screaming in their direction to stay away from the gate.[84]

This situation should not pose any real problem for human soldiers, who are likely to realise immediately that children chasing a ball do not represent a threat. Whether AWSs could reach the same conclusion is however unclear, since 'certain distinctions far surpass the abilities of today's robotics, at least at this stage: distinguishing a weapon from a cultural or religious symbol; distinguishing the agonized face of a person in fear for her

---

[78] U. Pagallo, op. cit., 59 (emphasis added). See also M. E. O'Connell, 'Seductive Drones: Learning from a Decade of Lethal Operations' (2011) 21 *Journal of Law, Information & Science* 116, 133 ff.

[79] 'Drones and democracy. Unmanned aerial vehicles are changing the democracy that uses them' (*The Economist*, 1 October 2010) available at www.economist.com/babbage/2010/10/01/drones-and-democracy.

[80] U. Pagallo, op. cit., 58 ff.

[81] *Ibidem* 60 ff.

[82] R. Crootof, op. cit., 1866–1868.

[83] See, for instance, N. Sharkey, 'Automating Warfare: Lessons Learned from the Drones' (2011) *Journal of Law, Information & Science* 140, 143–144.

[84] M. Wagner, op. cit., 22.

or his children from a threatening face; distinguishing children playing from threats'.[85] It seems unlikely that AWSs are capable of undertaking the highly context-dependent and essentially qualitative assessments that war situations often require.

A third concern about the use of killer robots relates to another basic principle of IHL, the principle of proportionality. According to the First Additional Protocol to the Geneva Conventions, the principle of proportionality is violated by an attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.[86] Such a qualitative exercise can hardly be carried out by AI systems.[87] Furthermore, by removing the human element from war, the use of killer robots can contribute to increase the number of deaths as there will be no room for those human feelings that play a role in war contexts (fear, compassion, etc).[88] Already in 2010, the special rapporteur on extrajudicial, summary or arbitrary executions, Philip Alston, singled out a similar problem with respect to the use of drones in war. He noted that drones may help developing what he called 'a "Playstation" mentality to killing'.[89] If drones are controlled remotely, it will be easier for the 'cubicle warriors' who 'operate from behind computer screens, physically far away from the battlefield'[90] to kill other persons than it would be for a soldier on the ground. The distance from the battlefield can become even greater when AWSs will be used; in this case, disincentives to kill can drastically decrease or even disappear. Furthermore, if AI systems are maliciously or improperly designed, their use can jeopardise the other fundamental principle of IHL according to which weapons that cause superfluous injury or unnecessary suffering shall be prohibited.[91]

Finally, it flows from the above that killer robots can easily end up committing international crimes. This brings the issue of AI systems' criminal liability back up. The discussions and outcomes of section I of the Congress should therefore inform also the last section, as the attribution of criminal responsibility for (international) crimes committed by robots is an unresolved matter under ICL as well. In this context, the issue is perhaps even more complex since liability for international crimes usually involves high level politicians or civil servants (doctrine of command responsibility), and their liability may

---

[85] *Ibidem* 23. See also P. Alston, op. cit., 54–55; T. Krupiy, 'Regulating a Game Changer: Using a Distributed Approach to Develop an Accountability Framework for Lethal Autonomous Weapon Systems' (2018) 50 *Georgetown Journal of International Law* 45, 48–50.

[86] Art 51(5)(b) of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.

[87] N. Sharkey, op. cit., 144–145; M. Wagner, op. cit., 23 ff.

[88] *Ibidem* 40 ff.

[89] Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Philip Alston (28 May 2010) A/HRC/14/24/Add.6, para 84.

[90] L. Royakkers and R. van Est, 'The cubicle warrior: the marionette of digitalized warfare' (2010) 12 *Ethics and Information Technology* 289, 291.

[91] Cfr. M. Hagger and T. McCormack, 'Regulating the Use of Unmanned Combat Vehicles: Are General Principles of International Humanitarian Law Sufficient?' (2011) 21 *Journal of Law, Information & Science* 74, 80–81.

not be easy to detect when it comes to crimes committed by means of killer robots.[92] The risk is to create a 'system of organized irresponsibility that shuffles responsibility from one actor to another without holding anyone accountable in the end'.[93] Some authors therefore suggest to use a 'distributed approach' to accountability, which 'ascribes responsibility to a senior political leader, a senior defense official responsible for promulgating policy on [lethal AWSs], a weapon manufacturer, a weapon designer, a military commander, and an operator'.[94] In practice, however, such a system may not play out well.

As is the case with cross-border cooperation on evidence gathering, therefore, an international approach on AWSs should be explored.[95] Some call for an absolute ban on the use of AWSs in war,[96] at least as long as their use is unlikely to be compliant with the core principles and rules of IHL,[97] while others support the conclusion of an international agreement that regulates the development and use of AWSs.[98] The XXI Congress will represent an invaluable occasion to discuss whether similar stances are feasible and really capable of reducing the significant risks connected with the use of killer robots.

[92] See, for instance, P. Alston, op. cit., 51–52; J. D. Ohlin, 'The Combatant's Stance: Autonomous Weapons on the Battlefield' (2016) 92 *International Law Studies* 1, 14 ff.; T. Krupiy, op. cit., 51 ff.

[93] M. Wagner, op. cit., 39.

[94] T. Krupiy, op. cit., 45.

[95] Cfr. M. Brundage et al., op. cit., 42.

[96] This position is supported by several NGOs, including ICRAC (International Committee for Robot Arms Control), which, together with other organisations, has launched the ongoing 'Campaign to Stop Killer Robots'.

[97] It has been claimed that '[AWS] should not be deployed at all until the deploying country, and by extension the international community, has satisfied itself that doing so can be done consistent with the requirements of international humanitarian law' (M Wagner, op. cit., 51). See also Human Rights Watch, op. cit.

[98] G. Bills, 'LAWS unto Themselves: Controlling the Development and Use of Lethal Autonomous Weapons Systems' (2015) 83 *The George Washington Law Review* 176.

RESOLUTIONS
of the XXIst International Congress of Penal Law,
Paris, 25-28 June 2024

RÉSOLUTIONS
du XXIe Congrès International de Droit Pénal,
Paris, 25-28 juin 2024

RESOLUCIONES
del XXI Congreso Internacional de Derecho Penal,
París, 25-28 junio 2024

XXIst INTERNATIONAL CONGRESS OF PENAL LAW
(Paris, 25-28 June 2024)

Topic: Artificial Intelligence and Criminal Justice

1. Traditional Criminal Law Categories and AI: Crisis or Palingenesis?

2. Penal Law and Criminalization in the face of the Challenges of AI

3. AI and the administration of criminal justice: 'Predictive policing,' 'predictive justice,' and evidence

4. International Perspectives on AI: Challenges for Judicial Cooperation and International Humanitarian/Criminal Law

Section I: Traditional Criminal Law Categories and AI: Crisis or Palingenesis?

Preamble

Considering that

- the advent of Artificial Intelligence (AI) systems with different degree of autonomy supports and replaces many human activities;

- AI systems might represent a real benefit for the society as a whole and for the law enforcement community, specifically when it comes to the investigations of criminal offences;

- AI systems are becoming increasingly autonomous, and their functioning may be unpredictable to those who design, program, produce, distribute and use them;

Observing that

- the areas of application of AI systems are considerably wide, and illicit facts related to their implementation might harm different interests, legal goods and fundamental rights;

- AI systems may also play an increasingly role in the perpetration of criminal acts as 'instrument' to commit criminal offences, and they might become the facilitator factor of the emergence of new criminal facts;

Paying particular attention to

- the increasing full or partial delegation of decisions to AI systems in different areas of activities, which raise the question of natural or legal persons' liability for the harms caused by the autonomous functioning of such systems;

- the autonomy of AI systems, that also created a debate within academia related to the possibility to consider them as the perpetrator of a crime;

Bearing in mind

- the importance of appropriate reactions that Criminal Law is required to provide in preventing and punishing offences committed by, through or against AI systems;

- the seriousness of the harms and of the risks related to AI applications;

- the fundamentals principles that must be guaranteed in establishing and applying criminal sanctions (including also punitive sanctions in a broader sense, which could be applied to legal persons), such as the principle of legality and the principle of culpability, which is a necessary expression of the personality of penal responsibility;

- aware that traditional Criminal Law categories and criminal responsibility models need to be considered and, if necessary, adapted to the emerging protection requirements;

Taking into account

- the 'Ethics Guidelines for trustworthy AI', presented to the European Commission on 8 April 2019 by the High-Level Expert Group, and other significant recommendations of other international bodies (for instance the 'Feasibility study on a future council of Europe instrument on Artificial Intelligence and Criminal Law' of the European committee on crime problems of the Council of Europe, 4 September 2020);

- the proposed European Regulation on Artificial Intelligence (so-called AI Act); the works of the Committee on Artificial Intelligence of the Council of Europe; the United Nations Activities on Artificial Intelligence;

- the recommendations of the XIV International Congress (Vienna, 1989), on the legal and practical problems posed by the difference between criminal law and administrative penal law, those of the XVIII International Congress (Istanbul, 2009), about the incrimination of preparation and participation in a crime, and those of the XIX International Congress (Rio de Janeiro, 2014), on Information Society and Penal Law,

The participants of the International Colloquium of the Section I, held in Siracusa (14-17 September 2022), have adopted what follows:

Recommendations

I. On the concept of Artificial Intelligence and the attribution of legal personality to AI systems with different degree of autonomy

1. Considering the increasing and complex evolution of AI, this concept encompasses multiple algorithmic and robotic systems that interact with the environment, developed with several techniques (such as Machine Learning), for the pursuing of human-defined objectives. Therefore, it is not desirable to provide a general AI definition for criminal law purposes.

2. However, since AI systems can be harmful in several fields (*e.g.* self-driving vehicles, robotic systems in medicine, trading AI systems or logistics management), the protection of legal goods and fundamental rights also under Criminal Law should take into account the specific features of the different AI systems with different degree of autonomy, as well as the legal definitions provided by extra-criminal law sources in each specific sectors.

3. As it stands, there is no normative ground nor consistency relating to the functions of criminal punishment in recognizing legal personhood to the Artificial Intelligence systems with different degree of autonomy.

4. On the one hand, there is an ontological distinction from human agents. AI systems lack the consciousness in choosing and evaluating the possible solutions to a problem or dilemma, considering also the context of social and ethical relationships and opportunities, with the necessary flexibility and capacity to adapt to even contingent or supervening situations and conditions.

5. On the other hand, punitive sanctions to such technological systems and agents would not respond to the purposes and functions of criminal punishment, because the effect of the threat of the penalty and its application would be emptied by the absence of self-awareness of their own existence in the past, present and future, and, above all, by the absence of voluntary self-determination, so that even excluding the retributive function, not even those of special and general prevention would be feasible.

II. On the need for extra-criminal regulation, standards and obligations

6. In order to prevent and reduce AI-related harms, before or, at least, in parallel with criminal law reforms, it would be necessary for International, regional, national legislators and competent authorities to fully define the regulation of the several sectors in which AI systems are implemented (such as those mentioned above of self-driving vehicles, health and surgical robots, autonomous weapons, etc.). Technical standards, structural characteristics and operating conditions of AI systems and their components should be regulated.

7. Such regulations, which must operate from the design, production, distribution and sales phases to the actual use of AI systems, should also provide for concrete requirements concerning adaptation in case of red flags or warning signals, as pre-condition for addressing AI-related harms through punitive law.

8. These regulations might provide for injunctional procedures, such as those already provided in areas of complex risks (*e.g.*, health and safety in the workplace and environment protection), the violation or non-compliance of which may be sanctioned according to the *ultima ratio* principle.

III. On the need for Criminal Protection of Legal Goods

9. It is necessary to recognize the essential importance of a reasonable and proportionate intervention of Criminal Law in a broad sense in preventing and punishing harms and dangers to interests, legal goods and fundamental rights that AI systems might cause, given that the same facts, if realized by natural and legal persons, according to the traditional categories of Criminal Law, might constitute a criminal offence. Therefore, they cannot go unpunished simply because they are carried out by or through or against the aforementioned systems.

10. It is necessary to identify and define specific models for attributing liability to the persons (both natural and legal persons) who are 'behind' the AI systems (i.e., the actors of the different phases of the life cycle of AI systems, such as designers, providers, importers, distributors, users, etc.), starting with the owners and those who decide on their concrete use, based on their interest and their benefit, and who must therefore be held legally liable, also from a 'punitive' – not only a Criminal Law - perspective.

11. The responsibility of the persons described in the previous point does not exclude that of other persons (either natural or legal persons) who contribute to the causal chain of the harm: from the designer, programmer, producer, seller, distributor to the end-users of the systems themselves.

12. In particular, a distinction must be made between:

a. AI systems used in illicit activities: in this area, there will mainly be malicious conducts, which pose fewer problems in terms of attribution of criminal liability, given that AI systems are conceptually no different from other instruments and means of committing a crime.

Two issues, however, should be addressed:

a.1. in case of deviant results of the functioning of the system from the intended illicit activity, the traditional principles of *aberratio ictus* and *aberratio delicti* must be applied. The mere material diversity of the harmed object must not represent an excuse if its characteristics are not relevant for the configuration of the criminal offence (*e.g.* killing one person instead of another is not relevant for the realization of the crime of murder, when it is intended by the agent). Instead it should be preferred to base criminal liability for a

crime other than the one intended on the possibility of concretely foreseeing such a different development of the action put in place by the AI system, by applying the principles of negligence based liability (as set out in chapter IV below);

a.2. since AI systems can be used for particularly harmful or dangerous conducts, they can amplify and aggravate the harm caused (as happen with ICT). Indeed, the consequences can be very distant from the actions that gave origin to them, making it more difficult to intervene *post factum* to prevent or at least to stop or reduce the harmful events. Therefore, the incrimination, as autonomous preparatory offences, of illegally designing, programming, producing, distributing, selling and purchasing of 'malicious' algorithms, software, and AI systems should be considered. This criminal policy should be limited to AI systems that pose high risks to certain significant legal goods (such as life, body, or liberty of other human beings) and only in case of clear, actual, present danger (on the conditions required in incriminating preparatory acts, see the resolution of the Section I of the XVIII AIDP Congress in Istanbul, 2009).

b. AI systems used in lawful activities: This case rises the most delicate issues in reference to the area of 'permitted risk', which should be delimitated through the hoped-for regulation of specific security obligations and precautionary rules to be applied to the activities of design, development, production, distribution, sale, as well as use, of AI systems. The adjustment of the models of criminal liability in this area must address the friction that can be created between forms of responsibility for negligent behaviour and the technical features of AI systems, namely: (1) their autonomy; (2) the concrete unpredictability of their decisions and functioning; (3) the opacity of their regulatory mechanisms; (4) the complexity of their programming, development, production, updating and maintenance process.

IV. On the adaptation of the models of attribution of liability to the features of Artificial Intelligence systems, specifically to their degree of autonomy

13. First of all, a distinction must be made, also according to the already recognized graduated automation and autonomy of AI applications in several areas, between the different levels of decision-making and operational autonomy of AI systems, which go from those where the 'automatic' functioning allows the human agent to have significant control over the system, to those that are truly 'autonomous', where human intervention can only be distant, in time and in space, from the functioning of the AI system, which 'decides' based on the information collected and on algorithms that adapt to its experience, so that there is a structural margin of unpredictability of the concrete outcomes.

14. In relation to the different types of AI systems, the definition of specific rules and standards of functioning, as foreshadowed in the proposal for a European regulation on Artificial Intelligence, is of fundamental importance (see chapter II).

15. The most pressing need for adaptation of the traditional categories of Criminal Law concerns the area of AI systems with a greater degree of autonomy, which are also the

result to which current technological development and experimentation in many fields is tending, so that they will undoubtedly be even more important in the near future.

16. Under this perspective, the field of corporate punitive liability might provide a useful reference, as well as the fields of product liability and liability for the protection of health and safety in the workplace.

17. In these legally regulated fields, often harmonized at a European level, existing principles might be extended, with the necessary adaptations, to AI-related crime regulation. Such regulated fields require the preventive assessment of the risks inherent in the specific activities performed, which have margins of permitted risk and correlated obligations of risk prevention and containment (see chapter II), with specific regard to the sources of dangers and harms.

18. Duties to act, especially in the case of red flags, are imposed to the relevant categories of persons (human beings), operating according to their respective competences, i.e. users and persons having the position of guarantee. They must promptly adapt the regulatory and security measures of their activity, to the point of stopping it, if necessary.

19. From these recognized principles, the following recommendations can be elaborated to structure criminal liability for AI-related harms:

i. Criminal liability of natural persons. It must be based on the identification of personal positions of guarantee, in relation to the competences and functions performed in using AI systems. Firstly, it will have to be considered the actors participating to the different phases of use of AI systems, until the end-users. Secondly, the positions of top and middle management and compliance officers in complex organizations should be considered. In each case, the formalization of positive obligations, of a technical, organizational and control nature, shall be addressed.

Criminal liability for negligent behaviours must comply with the general principles of Criminal Law, namely, the principle of personal culpability, since the objective connection between the causal contribution of the human agent and the commission of the offence by the AI system does not suffice, given that the foreseeability and evitability of the illicit fact are also necessary. Particularly, criminal responsibility for negligence, for not having acted differently from what would have been possible, must be correlated not so much to the specific and concrete event or fact that occurred, as to the scheme of 'organizational fault', referring to the way the artificial agent is structured and operates. The assessment of the risks arising from the AI system's activities must also include the awareness of outcomes that are concretely 'unforeseeable' in individual cases, which is the basis of the obligation to prepare adequate and always up-to-date surveillance and containment measures, for which the natural person in charge remains responsible, being accountable (accountability), as the owner or top representative of the organization that uses the AI system in its own interest or to its own advantage.

ii. Punitive liability of legal persons. Considering that a large part of AI systems is produced or used by legal persons, it is of necessary to hold them accountable for the offences committed by, through or against such systems.

In this respect, assuming that precise public standards of conduct and compliance are to be introduced (cf. chapter II), punishment of the legal person, proportionate to the offences committed by, through or against AI systems and to the degree of fault of the organization, might be related to a model of liability based on organizational fault. Such model of liability leads to imputing responsibility subjectively, as the object of culpable reprehensibility, to the legal person in case of offences caused by the lack, deficiency or inadequacy of organizational and prevention measures, to be implemented and updated on the basis of the assessment of the specific risks deriving from the activities entrusted to and, in any case, carried out by the AI systems, in their interest or to their advantage. Strict liability models should be avoided.

A new model of corporate autonomous punitive liability, not based on the liability of the individual natural person, should be promoted, since the legal person can be held liable even if the natural person who realized the harm is not individually punishable due to particular conditions or circumstances or if he/she is not specifically identified. Indeed, it is enough to ascertain the commission of an objectively typical and unlawful act in the interest or to the advantage of the organization.

In those national legal systems that base corporate liability to a closed list of offences, it is to recommend the extension of such list to the criminal offences that can be committed through, by or against AI systems.

V. On preventive measures and punitive sanctions applicable to natural and legal persons 'behind' Artificial Intelligence systems

20. The sanctions applicable to natural persons, including imprisonment, and to legal persons, possibly of an administrative nature, according to the various legal systems, but in any case of a punitive nature, including fines and suspension of the activity by which the offence was committed, should in principle correspond to those applied for the type of offence realized, in accordance with the principles of the single legal systems, namely the principles of proportionality and individualization of the sanctions. When it comes to legal persons, these sanctions, in addition to pecuniary measures, should also include the injunction to modify the corporation's compliance and internal control system, as well as the possibility of ordering a period of public monitoring of the corporation to ensure that it complies with the imposed standards.

21. Given the seriousness of AI-related harms, it is to recommend the adoption of preventive measures for their effectiveness in averting or mitigating the harmful consequences of AI systems (as, *inter alia,* seizure, confiscation, judicial monitoring, interdictory measures).

22. The important role of non-pecuniary penalties, such as the sanction of disqualification from exercising specific activities and confiscation, should be emphasized. Specifically, confiscation allows direct action to be taken against the AI system with or by which the offence was committed, without the need to recognize it as a legal entity or as having criminal capacity (see chapter I above).

VI. Complementary enforcement systems

23. Given the problematic and foreseeable difficulty of implementing an effective criminal liability system for natural persons and legal entities 'behind' AI systems for offences committed by, through or against them, the adoption of complementary enforcement systems is to be recommended.

24. Such enforcement system might include administrative authorizations and certifications, as well as civil remedies.

25. Further alternative to criminal prosecution and/or sanctioning might include models of compliance, restorative justice interventions and agreements with victims and competent public authorities.

Section II: Penal Law and Criminalization in the face of the Challenges of AI

Preamble

Considering that

- Artificial Intelligence, one of the last advancements within the digital revolution, has already reached a significant level of development in this third decade of the 21st century and is already widely used in many sectors of society;
- Although there is no absolute agreement on the definition of this technology, there is an implicit consensus that it includes a multitude of computerized systems that, through the gathering, the processing, and the analysis of data in their context, are capable of acting autonomously and/or assisting in the decisions to achieve specific objectives.
- The transformative potential of this technology is impacting multiple fields and social spheres, bringing important benefits and opportunities,
- At the same time Artificial Intelligence poses risks and harms to individual and collective interests,

Observing that

- multiple decisions traditionally adopted or informed by humans are beginning to be automated through the use of these technologies, affecting different areas and interests;
- In areas such as autonomous vehicles, health services, financial markets, media, and

other sectors, the use of this technology is unstoppable and a future without it seems inevitable;

- The promise of efficiency and objectivity of AI is leading to the development of these technologies without assessing its real necessity or considering the risks that it may create.

Bearing in mind

- that recent developments in large language models and other AI systems such as machine learning and deep learning, have highlighted the need for regulation, including security protocols, to control the evolution of these technologies in terms of their effects and risks;

- that the development of AI systems, particularly the training of its algorithms, requires the use and accumulation of data and large amounts of information, which is a risk that must be considered in itself;

Highlighting

- the growing concern about the harmfulness of some malicious or negligent uses of AI can cause in areas where AI is already beginning to have a strong presence;

- that there are many countries in which the use of AI systems has caused harm to relevant interests such as life, health, privacy and others;

Acknowledging

- that the emergence of new criminal acts as well as new interests worthy of criminal protection will lead states to adapt criminal laws to the new realities related to AI,

- the need to analyse whether the legal response of states to the challenges of AI is sufficient or whether it needs to be reformed and adapted, either through specific modifications or through the creation of new forms of criminalization,

Taking into account

- the Recommendation of the Council on Artificial Intelligence, adopted by the OCD on the 22 of May 2019, the Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts presented by the European Commission on the 21 of April of 2021, the European Parliament resolution of 6 October 2021 on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters (2020/2016(INI)), the Ethics Guidelines for trustworthy AI presented by the High-Level Expert Group set by the European Commission on 8 April 2019 and Recommendation CM/Rec (2020) 1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on

8 April 2020 at the 1373rd meeting of the Ministers' Deputies),

- the debates and resolutions of previous International Criminal Law Congresses, in particular, the resolutions of the XIV Congress on Criminal law and modern bio-medical techniques and the legal and practical problems posed by the difference between criminal law and administrative penal law, the resolutions of the XIX International Congress of Penal Law - Information Society and Penal Law, and the draft resolutions of the XXI International Congress of penal Law on Section I (Criminal Law- general part) and Section 3 (AI and the administration of criminal justice),

The participants of the International Colloquium of the Section II, held in Bucharest (14-16 June 2023), have adopted what follows:

Resolutions

I. AI regulation and enforcement

1. With the development and expansion of AI, harms and risks to individual and collective interests have arisen and are likely to increase in the future. This poses the need for an adequate regulation of the production and implementation of AI systems and of its use. The response to these challenges shall consider different perspectives and all available regulatory tools, whether they are public or private, taking into account the different nature and functionality of each of them.

2. The global impact of AI requires an international response to reach an effective protection of individual and collective interests at stake. States should take into account international standards in national regulation and enforcement.

3. As happened with other technological and socio economical innovations, as for instance, the Internet or the new developments on gene editing and neuroscience, the irruption of AI makes necessary to review general aspects of the criminal justice system and, in particular, the catalogue of existing offenses in criminal laws, checking whether the regulation in national states is suitable to face the challenges posed by the use of this technology.

4. The debates regarding the transformation of Criminal law due to the impact of AI and the role that the criminal law system might play in relation to these new technologies shall not be isolated from the international ethical discussions that are taking place on the development of AI. Nor can they overlook the discussions regarding AI regulation on other branches of the legal system. Considering this broad picture, legislators shall reflect on the specific role that Criminal law is playing to avoid harms caused by the use of AI.

5. Legislators will have to reform existing offenses when AI modifies the dimension of

the risk to existing interests or creates new means of perpetration that are not covered by the existing legislation. Moreover, new offenses shall be introduced when the development of this technology leads to the emergence of new individual and collective interests worthy of protection not covered by existing legislation.

II. Criminalization and the protection of interests related to AI

6. The development of AI may give rise to new interests worthy of protection. Additionally, AI systems can affect the dimension and relevance of interests that are not currently considered worthy of protection by Criminal law. When criminal laws do not provide an adequate response to protect these interests, new criminal offenses shall be enacted that proportionally punish conducts that are harmful to such interests. This shall only be done when there are no alternative means that are less harmful than Criminal law to effectively protect the mentioned interests.

7. When the transformation of AI leads to the emergence of new interests, which are essentially similar to others traditionally considered worthy of protection, new criminal offenses shall not be included. Instead, it is preferred to adapt the interpretation of the existing offenses, as long as the strict respect for the principle of legality allows it.

8. Legislators will have to decide whether the development of this technology gives rise to the need for specific criminal protection of individual or collective interests related to the AI technology itself. While it is still too early to determine whether such a need will arise, this could be the case for the data on which the algorithms are based; the functionality of AI systems themselves in some cases; collective interests related to the safety and reliability of their design and application; or even interests associated with robots.

9. Some AI systems, such as those used in critical infrastructures, are essential to assure already protected interests. As Budapest convention contemplates the criminalization of attacks to computer systems, and AI may be considered as that, the enactment of new offenses might not be necessary. In order not to raise interpretative doubts, it might be advisable to reform some criminal offenses to introduce AI systems as types of computer systems.

10. As long as it is not proven that there are some other interests at stake and taking into account the current level of development of these technologies, AI and robotic systems do not deserve a different protection, in relation to their economic or functional value, than other computer systems.

III. Grounds of Legitimization and techniques of criminalization

11. Criminal law shall not play a leading role in the regulation of AI. In view of its nature as a particularly coercive enforcement instrument, it must intervene as the last resort and

limited to the repression of the most serious and harmful acts.

12. Legislators shall not introduce new offenses based solely on the fact that AI was employed. Many of the criminal offenses can be committed by using AI systems as a means of carrying out the sanctioned conduct. It is only when acts committed with AI systems acquire a different meaning in terms of harmfulness or risk, that it will be necessary to enact new criminal offenses.

13. The automation of data-driven decision-making processes that AI entails resituates the key moment of human agency to phases of design and implementation of algorithms far removed from the harm. It will therefore be in these phases where the liability of individuals and legal persons involved should preferably be focused. This shall be done considering existing legal duties established in other branches of the legal system.

14. Criminal law systems are designed to have a deterrent effect on likely offenders, preventing them from engaging in criminal actions. If the key moment in terms of risk in relation to AI is the moment of the design and implementation, the enactment of offenses that aim to deter conducts at such moments shall be considered. This can be done by anticipating the protection with endangerment offenses, that punish not following certain duties in relation to specific interests worthy of protection. Also, and similar to what is established with the criminal liability of legal persons, specific regulatory obligations related to the design and implementation of AI systems could be established, which infringement may give rise to criminal liability.

15. Endangerment offenses related to the design and implementation of AI systems that cause risks shall be enacted when the sanctioned actions pose a considerable threat to the protected interests. Additionally, the legal consequence attached to these offenses shall be proportional to the level of risk caused and the interest that is at stake. Due to the complexity in the design of AI systems and the different approaches to the regulation of these tools, endangerment offenses shall not be enacted before considering the developments in self-regulation or administrative regulations on control and security of AI in each legal system. These regulations shall serve to identify relevant risky acts that might be worthy of criminal prosecution.

16. New offenses might be introduced to punish the abuse and transformation of existing lawful AI systems when, by changing the design or the purpose of the AI, new risks arise.

17. In those legal systems where negligent action is only punished when expressly provided for (*numerus clausus*), the reform of criminal laws might be required. The complex design of AI and the participation of multiple parties in the AI lifecycle means that in most cases it will be extremely difficult to prove awareness that a harmful result was going to occur at the time of design. Negligence offenses based on the infringement of standards of due diligence could be then enacted if the protection of the affected

interests make it necessary.

18. Since AI systems are dynamic and their performance depends on the introduction or collection of data that modify its outcomes, risk management processes established on other branches of the legal system might operate throughout the entire AI lifecycle. Thus, criminal laws may address, if necessary, infringements of rules related to the lack of appropriate monitoring and oversight of AI systems, duties that might affect different subjects involved in the whole lifecycle of AI.

19. Since AI technology and its applications are scalable, criminal justice systems might adapt to adjust the proportionality of penalties to the severity of the harm that AI can cause. Nevertheless, legislators shall not enact aggravated circumstances only because an offense was committed using AI. Only if existing aggravated circumstances can't encompass the severity of the damages caused by the use of AI, attending also at the relevance of the affected interest, new forms of aggravation shall be considered. This shall always be done complying with the principle of proportionality.

IV. Criminalization and the protection of specific interests from the risks created by AI.

20. Since criminal laws do not usually provide for specific means for the commission of crimes against life and health, it does not seem necessary to reform these offenses in order to protect such interests when AI system have been used as a mean of commission. Nor does it seem necessary to modify the system of liability graduation. Nevertheless, in specific areas, such as autonomous driving, criminal law will have to be attentive to the changing regulatory landscape, that will be the basis to assert what is considered an 'allowed risk' for the determination of liability.

21. If autonomous driving becomes widespread, road safety offenses could undergo significant changes, including new offences related to new risky behaviors for life and road safety other than those currently focused on human driving.

22. In the same way that the revolution on gene editing led to the appearance of offenses sanctioning genetic manipulation with the capacity for mass destruction, technological evolution may make it necessary, in the near future, to criminalize the creation, development and use of AI tools with a high destructive capacity, such us some autonomous weapons, drones or robots that could be enormously harmful specially if human control is lost.

23. AI systems collect and rely on large amounts of information to perform its tasks, creating new threats to classical interests. Given this development of AI technology, a review of offenses linked to privacy and other personal interests is necessary, and a revision of the understanding of privacy as a solely individual good, considering a collective dimension of this interest, should be taken into account.

24. The criminalization of acts involving the unlawful gathering of personal data should not only be linked to the protection of interests such as privacy. The use of AI in cyberspace may open the door to mass data collection for the commission of cybercrimes that harm interests such as property. States should review whether it is necessary to enact criminal offenses to sanction the unlawful massive collection of data, and similar preparatory acts to serious crimes, whenever it causes concrete risk to those interests and only if there is not another less coercive legal tool available.

25. The accessibility of images and personal data in cyberspace linked to the potential of generative AI to transform images, video, and audio, can endanger interests such as reputation and honor or sexual freedom. It is necessary to review whether current criminal laws allow the punishment of harmful conducts to human dignity, reputation, and sexual freedom such as the distribution of Deep Fakes, including those with sexual content, or of child pornography.

26. Generative AIs, such as Large Language models and other similar tools, can facilitate deception, threats, and coercion, affecting different phases of the formation of will, endangering interests worthy of protection. Nevertheless, it does not seem appropriate to introduce new criminal offenses, since current criminal laws encompass the more harmful acts and other means of controlling this type of tools should be used to prevent less risky conducts.

27. The popularization of algorithms for risk management in areas as healthcare, employee recruitment, justice, credit and loans and many others, has revealed the existence of discriminatory biases in some of these decisions taken by AI systems. Beyond the criminal offenses that can already punish some particularly serious discriminatory decisions, other branches of the legal system such as civil or administrative law are more appropriate for avoiding the problem of algorithmic discrimination.

28. The use of AI in cyberspace may facilitate and enhance existing attacks against property and other interests worthy of protection. However, given the regulation of cyber fraud and other cybercrimes against property, it will not be necessary, at least in the short term, to adapt these offences to accommodate crimes perpetrated using AI systems.

29. Some of the criminal laws which punish the production, sale, procurement for use, import, distribution or otherwise making available of devices designed or adapted primarily for the purpose of committing offenses enacted in accordance with articles 2 through 5 of the Budapest Convention can already punish the creation, development and sale of AI systems designed or adapted for those criminal purposes. Thus, as long as AI is considered a device, including a computer program, in accordance with Art. 6 of the Budapest Convention, the introduction of new offences that anticipate the criminal response is not necessary in this field.

30. In the socio-economic and financial sphere, the proliferation of algorithmic decision-making systems and the use of AI for trading is already being reported. The risk that malicious or negligent use of AI systems could seriously affect the markets is clear, but the interests at stake could be better protected through preventive regulatory measures of an economic, administrative and commercial nature, rather than through new criminal offences other than those already in place to punish insider trading and other similar conducts. When AI systems are used for manipulating markets criminal law should be revised to give a proportional response.

31. Concern about the impact of the phenomenon of disinformation, first in the aftermath of some electoral processes and then with the infodemic during the Covid-19 crisis, has led many states to create criminal offenses to punish this conduct. The fact that AI can increase its impact either by automating its dissemination or by using sophisticated video, audio, image and text manipulation technologies may sustain this trend. Criminalizing disinformation will only be justified for the protection of fundamental interests of democratic societies and if does not jeopardize freedom of expression.

Section III: AI and the administration of criminal justice: 'Predictive policing,' 'predictive justice,' and evidence

Preamble

Aware that artificial intelligence (AI) is developing rapidly in contemporary society in various parts of the world. Already ubiquitous in peoples' lives in some countries, it may become part of daily life for a large part of the world's population in the future.

Noting that as a technological innovation, AI pushes consumers to buy new products, thus helping the global economy grow. Therefore, AI plays a non-negligible role in sustaining and even expanding the liberal market economy and the capitalist economic system.

Noting that the companies that create and market AI are frequently based in developed countries in the global North, and they often try to open markets all over the world.

Noting that 'digital divide' widens social inequalities among people. 'AI divide' may be the next phenomenon on the horizon.

***

Considering that AI can be defined as a set of theories and techniques used to create machines capable of simulating human intelligence.[1] As a scientific discipline, it is a blend of statistical and algorithmic mathematics, computer science, and the cognitive sciences. Symbolic AI is based on the rules of logic, whereas connectionist AI uses artificial neural networks.

[1] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

Considering that machine learning is an example of connectionist AI, as is deep learning, which is a subset of machine learning that uses multiple layers of interconnected artificial neurons. As the number of neuronal layers enabling autonomous learning increase, the system's technological complexity increases, making the system more efficient and its calculations less explainable and traceable (deep learning).

Considering that machine learning can make technology extremely powerful, but its decision-making process can be so complex that it resembles a 'black box'.

Noting that many AI systems used in the context of preventing, investigating, detecting, and punishing crime are machine-learning systems. Using self-learning algorithms, they carry out complex probability calculations in nanoseconds. To achieve their assigned goals, they process huge amounts of data and consume a lot of energy. Some of them, such as facial recognition systems, rely on deep learning.

***

Finding that in criminal justice administration, AI systems are used to prevent or detect criminal offenses based on risk assessment ('predictive policing').

Finding that AI systems are also used to help prosecutors and judges make decisions. More specifically, the term 'predictive justice' refers to (i) anticipating someone's behavior, *e.g.*, to assessing the risk of fleeing in the pre-trial procedure or (re-)committing a crime so that decisions concerning them, such as pre-trial detention, sentencing, parole, and probation (actuarial justice, that nowadays may be supported by AI) may be made; and (ii) using AI to perform an ultraquick statistical analysis of prior decisions issued in similar cases and of relevant legal and regulatory provisions (quantitative legal analysis or LegalTech).

Considering that quantitative legal analysis is revolutionary in the sense that a mathematical calculation is meant to support or even to supplant legal reasoning.

Observing that the word 'predictive' used in the phrases 'predictive policing' and 'predictive justice' is confusing because AI systems calculate probabilities, but do not predict the future; these probabilities are based on correlations, not on causations. These calculations nevertheless have a performative effect on people, that is, might induce them to decide in line with their results. General speaking, AI's scientific roots encourage its users to trust and follow the probabilities calculated by the AI system, since 'automation bias' is higher when the system embodies a degree of scientific aura.

Finding that AI systems contribute to innovation in the search for evidence. They can quickly analyze big data and extract information that can be useful to investigators. AI systems can establish correlations between pieces of information that are invisible to the human eye. The crime analysis diagrams they produce can be highly valuable, elaborate information for investigators.

Finding that AI systems can produce information proffered as evidence for use in criminal

trials. In particular, AI systems can provide forensic information by comparing biometric traits (*e.g.*, facial images in facial recognition), the sound frequencies of different voices (vocal recognition), and DNA fragments (probabilistic genotyping).

Finding, lastly, that AI-assisted robots and 'smart' objects in various occupational areas and daily life may incidentally produce clues or evidence that may turn out to be useful in establishing facts in a criminal case.

***

Finding that despite significant progress in the past few years, AI systems are not completely reliable. Errors may be due to the poor quality of the data used or to how the algorithm is programmed or to the existence of false positives/negatives in correlations. The probabilities produced by an AI system may therefore be inaccurate.

Finding that the results produced by AI systems are not always entirely neutral. The accuracy of the probabilities calculated by AI systems depends not only on the quality of the collected and processed data, which may reflect bias, but also depends on how the systems have learned (unsupervised vs. supervised learning). Because they reproduce human decisions, self-learning algorithms are influenced by human foibles. One result is xenophobic,[2] racist, misogynist, etc. algorithms.

Finding that AI systems pose transparency problems. So-called 'black box AI' is so opaque that even specialists cannot determine how it arrives at its results. Even scientific experts cannot fully explain a system's reasoning to a court.

Considering that AI systems used in the field of criminal justice may be developed by the private sector. Such systems are products to be sold and must be profitable. The companies that develop them generally invoke trade-secret protection to refuse to reveal their algorithm's source code, without which the system's functioning cannot be properly analyzed.

***

Observing that not everything that is technologically possible is socially desirable. In a democracy, the political choices affecting the prevention, detection, investigation, and punishment of criminal offenses must be reflected in a law or a norm of equivalent binding force.

Reiterating that human rights must be fully protected when preventing, detecting, investigating, and punishing offenses, including when technological innovations are used in that context. Whereas AI often raises issues of privacy and personal data-protection law as well as the law of non-discrimination, all laws protecting human beings, in particular

[2] https://www.amnesty.org/en/documents/eur35/4686/2021/en/

their freedom and dignity, as well as all the guarantees of a fair trial, including the presumption of innocence, are potentially threatened by the use of technologies that simulate human intelligence.

Considering that national laws and international and/or regional legal norms can set out the terms on which AI-related technological innovations may be allowed to contribute to the administration of criminal justice.

Reiterating that the ethical standards often referred to by the private sector do not have the same binding force as law.

***

Aware of:

- the Recommendation of the council on artificial intelligence, Organization for Economic Co-operation and Development, 22 May 2019, C/MIN(2019)3/FINAL;

- the Recommendation on the Ethics of Artificial Intelligence, United Nations Educational, Scientific and Cultural Organization, 23 November 2021, SHS/BIO/PI/2021/1;

- the European ethical Charter on the use of Artificial intelligence in judicial systems and their environment, European Commission for the Efficiency of Justice, Strasbourg, 3-4 December 2018;

- the European Parliament Resolution of 6 October 2021 on artificial intelligence and criminal law and its use by the police and judicial authorities in criminal matters, document 2020, 2016 (INI);

- the resolutions of the XIXth International Congress of Penal Law: Information society and penal law, Rio de Janeiro, 2014.

***

The participants of the International Colloquium of the Section III, held in Buenos Aires (28-31 March 2023), have adopted what follows:

Resolutions

1. Use of AI systems by public authorities for assistance when preventing, detecting, or investigating criminal offences must be authorized in advance by a law or a norm of equivalent binding force.[3]

2. States must ensure that the decisions taken by authorities to focus on preventing, detecting, or investigating a particular type of crime is based on politically and democratically determined criteria rather than on the assumption that using AI technology will

[3] Below we will shorten 'norm of equivalent binding force' to 'equivalent norm.'

make it easier to prevent, detecting or investigating this type of crime.

3. To protect the legitimacy of the public authorities' activities preventing, detecting, and investigating criminal offences, states that wish to use AI systems must choose systems the functioning of which is fully transparent, explainable, and traceable (white box AI). They must ensure that intellectual property objections cannot be raised when seeking transparency, and they should prefer publicly available, open-source systems.

4. Laws or equivalent norms related to using AI systems in the prevention, detection, and investigation of criminal offenses must require that such systems have a high degree of technological reliability. A sufficiently precise regulation requiring appropriate verifications and evaluations, both external to and independent of the AI system's developer and provider, must limit to the greatest possible extent the risk of bias or any form of discrimination in machine learning, coding errors, and other technological malfunctions.

5. Laws or equivalent norms must require that AI systems used to assist in the prevention, detection, and investigation of criminal offenses be fully accessible, verifiable, and auditable by authorities that use them and by authorities that are in charge of verifications and evaluations.

6. Laws or equivalent norms authorizing the use of AI systems to assist in the prevention, detection, and investigation of criminal offenses must require that the training data be of high quality and representativeness.

Concerning data from police or judicial files, laws or equivalent norms must institute a system that ensures that such data are correct and up-to-date and that their use does not infringe the presumption of innocence. The presumption of innocence strictly prohibits the retention and use of data collected in response to the outcome of a predictive assessment when there is no subsequent finding of guilt, except if the data have relevance concerning another suspect.

As regards other data, in particular data accessible on social media, laws or equivalent norms must require compliance with the right to privacy and with personal data protection law when using such data. Appropriate verifications, independent of the police and judicial institutions, must be undertaken.

In general, laws or equivalent norms must be highly demanding with respect to the verification of the reliability of all data used by AI systems in connection with detecting, preventing, and investigating criminal offenses.

7. Laws or equivalent norms must require that before an AI system based on self-learning algorithms may be used in preventing, detecting, or investigating criminal offenses, the algorithms must be developed, trained, tested, and deployed under human supervision (human-in-the-loop machine learning).

These laws and equivalent norms must require a human evaluation before any action is taken to prevent, detect, or investigate criminal offenses based on the probabilities calculated by an AI system.

8. States and law enforcement authorities must ensure that their personnel who use AI to prevent, detect, or investigate criminal offenses receive hands-on training in the proper use of the relevant AI system, as well as training with respect to the risk of error and bias. They must ensure that such personnel have a thorough knowledge of the dangers AI may pose to human rights.

9. International, regional, national, or local authorities must establish independent bodies certifying the quality of AI systems intended to be used in preventing, detecting, or investigating criminal offenses. AI technology that cannot be operated and supervised in a transparent way, due to, *inter alia*, intellectual property rights, must not be certified.

The private sector should organize or unite to create AI-system quality labels with the goal of creating a virtuous circle for these products so that the authorities working to prevent, detect, or investigate criminal offenses are better able to determine which AI systems meet their needs.

10. All human rights must be protected when AI systems are used in preventing, detecting, or investigating criminal offenses. States and regional and international bodies must ensure that effective, proportionate, and dissuasive sanctions are imposed when such rights are violated.

Laws or equivalent norms must explicitly provide that where the cause of the violation of human rights is the technological malfunction of an AI system, the company that created the system will incur liability for fault or negligence or based on strict liability for defective products. They must also provide that investigations must be carried out to determine the cause of the violation.

11. All present resolutions are also applicable to preventing, detecting, investigating, and sanctioning administrative offenses by the competent authorities.

Resolutions specific to 'predictive policing'

12. States and regional and international human rights bodies must ensure that the use of AI systems in preventing and detecting criminal offenses does not lead to mass surveillance, which would result in a disproportionate reduction of individual freedoms (freedom of movement, freedom of expression, freedom of assembly, freedom of association, and freedom of religion).

In particular, states and local authorities must prohibit the use of AI systems to remotely identify individuals in publicly accessible spaces on the basis of their biometric data, as well as any other uses of AI systems that enable mass surveillance.

States are urged to be more transparent about their use of automated number plate recognition systems in publicly accessible space. When these systems include not only the taking of a picture of the licence plate, but also the taking of a picture of any individual in the vehicle, this option must be explicitly authorized by law. Applying facial recognition technology to the data collected through these pictures must be prohibited for the purposes of 'predictive policing'. It can only happen in the context of a specific investigation if there is a legal framework for it.

13. States must determine or have independent research bodies determine whether using AI systems in preventing criminal offenses helps decrease the number of offenses committed and, if so, in what proportion.

14. States must ensure that the financial cost of AI systems and their maintenance does not deprive the public crime-prevention services working on the *causes* of crime of funds (for psychological support, social support, training, and employment support).

15. Laws and equivalent norms must strictly prohibit the use of data as inculpatory evidence in criminal proceedings where those data were collected by an AI system in connection with crime prevention, that is, where there was no concrete suspicion that an offense had been committed and therefore the data were collected outside the scope of the legal framework governing criminal investigations.

If data collected by an AI system in the context of crime prevention are used as the basis for investigation ('starting information), in criminal investigation as starting information, the competent judicial authority must be informed of it. The data must be marked as such and the use of AI systems must be documented on the case file.

Resolutions specific to 'predictive justice'

16. Laws and equivalent norms must strictly prohibit the use of AI systems for actuarial justice purposes in sentencing.

Punishing or aggravating the punishment of someone based on the probability that they will commit a criminal offense in the future amounts to applying punishment based in part on a criminal act that has not occurred. That is contrary to human dignity, personal freedom, and fundamental principles of criminal justice.

The use of AI risk-assessment tools must be prohibited when severe security measures, such as detention, come into consideration. When states allow the use of such tools for less severe measures, the law must expressly authorize it, with sufficient procedural safeguard. However, AI probabilities cannot constitute the only basis for a decision.

17. States that wish to use AI to assist prosecutors or/and judges with quantitative legal analysis before taking decisions in criminal cases must limit use of this technology to minor offenses that represent a high volume of cases.

18. Before deciding to use AI to facilitate management of a high volume of cases involving minor offenses, states must assess whether it would be appropriate, in light of the *ultima ratio* principle, to decriminalize the conduct generating such cases.

19. Laws and equivalent norms must prohibit the use of quantitative legal analysis for assisting judges when ruling on guilt.

20. Laws and equivalent norms must prohibit the use of quantitative legal analysis for assisting judges with sentencing. The decision to punish a person and the type of sentence must be made by humans. Otherwise, justice may be dehumanized and people's human dignity may be threatened.

21. Laws and equivalent norms must prohibit the use of quantitative legal analysis for assisting judges with decisions in criminal matters that are issued before judgment and that involve coercive measures.

22. States must ensure that decisions taken with the assistance of quantitative legal analysis do not infringe the right of access to a human judge.

23. Laws and equivalent norms must prohibit the assistance of quantitative legal analysis unless the decision can be appealed by the person concerned. The decision at appeal level shall not be based solely on the quantitative legal analysis.

Resolutions specific to evidence gathered and/or produced by AI

24. Laws and equivalent norms on extracting data for analysis by an AI system must require that before asking a person for the access code of her/his software or hardware from which data may be extracted, the seizing authority must inform the person concerned of their right not to incriminate themselves.

25. Laws and equivalent norms on crime analysis must specify that the crime analysis diagrams produced by AI systems do not have probative value, but may serve as a guide for conducting investigation.

26. Laws and equivalent norms on using AI systems to gather evidence or produce information for criminal justice purposes must clearly indicate that the output of AI systems are only probabilities. They must require that all probability-based judgments indicate not only the probability calculated by the AI system that was used, but also the error rate of that system, as calculated by the certification body that evaluated it.

27. States and judicial authorities must ensure that the use of AI-calculated probabilities does not lower the existing standard of proof in criminal proceedings.

28. Laws and equivalent norms on using AI systems to gather evidence or produce information for criminal justice purposes must prohibit the use, as evidence, of probabilities calculated by AI systems that are not fully explainable (black box AI).

29. Laws and equivalent norms on using AI systems to gather evidence or produce information for criminal justice purposes must require, pursuant to the right to adversarial hearings, that, if data collected or produced by an AI system are used, all parties must be informed of it. The data must be marked as such and the use of AI systems must be documented on the case file.

Laws and equivalent norms must require that a party's production of an AI-calculated probability may be challenged by the other party.

30. Laws and equivalent norms must set forth the principle that the party producing the probability in court must systematically include complete information on how the AI system works and which data it uses.

31. Laws and equivalent norms on using AI systems to gather evidence or produce information for criminal justice purposes must, consistent with defense rights, provide that anyone accused of an offense based on a probability proffered as evidence be able to obtain the AI system's source code and training data so that these may be analyzed by an expert. Trade secret must not be allowed to impinge on defense rights.

32. Due to the high cost of obtaining an expert analysis of an AI system, states must ensure that anyone accused of an offense based on a probability calculated by an AI system have access not only to effective legal aid but also to financial aid for such specific expertise.

Section IV: International Perspectives on AI: Challenges for Judicial Cooperation and International Humanitarian/Criminal Law

Preamble

Considering that the artificial intelligence (AI) is an integral part of the Autonomous Weapons Systems (AWS) decision-making process, and such weapons have already reached a significant level of development in this third decade of the 21st century and that they are already widely used by many states and non-state actors.

Considering that multiple surveillance and targeting decisions traditionally adopted or informed by humans are beginning to be automated through the use of AWS, causing significant legal, moral, and ethical considerations;

Highlighting the growing concern about the harmfulness of malicious or negligent uses of AWS;

Acknowledging the need to analyze whether the legal response of states and the international community to the challenges of AI is sufficient or whether it needs to be reformed and adapted, either through specific modifications or through the creation of new forms of criminalization;

Acknowledging the need for AWS to be designed and developed in compliance with the international law.

***

Noting that AWS can be defined as '[a]ny weapon system with autonomy in its critical functions—that is, a weapon system that can select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention;'[4]

Noting however that this definition of AWS is not universally accepted, and that different international, regional and national regimes may choose to adopt different understandings of the definition of AWS and of meaningful human control in their use ('human on the loop');

Noting that different international, regional and national regimes may also choose to distinguish between AWS and Lethal Autonomous Weapons Systems, and also between AWS and Automated Weapons Systems;

Considering that the use of AWSs may cause significant harm to humans and societies, and that such use raises legal and ethical implications related to both *ius ad bellum* and *ius in bello*, serious human rights violations, *ius cogens* violations and core international crimes;

Considering that the use of AWSs can influence public opinion and policy in favor of the use of force, because of the perception that the use of AWSs minimizes risks of destruction, death or bodily injury to soldiers and other individuals involved;

Considering that the use of AWSs can negatively affect the respect of fundamental principles of *ius in bello*, such as the principles of distinction, proportionality and precaution(s);

Considering that by removing or reducing the human element in the decision making, the use of AWSs can contribute to the increase in the number of deaths because of the absence of human feelings, such as fear and compassion, which may play a role in reducing the number of deaths;

---

[4] ICRC, *Views of the ICRC on autonomous weapon systems*, paper submitted to the Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 11 April 2016, https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system

Considering that the use of AWSs may cause significant destruction and collateral damage;

Considering that core international crimes can be committed through the use of AWSs and that such use of AWSs may raises serious and new issues related to the attribution of criminal responsibility questions, including, but not limited to issues related to command responsibility.

***

Considering that the use of AWSs can raise jurisdictional issues, because AWSs use may be trans-territorial;

Finding that an international regulatory approach to AWSs is necessary and should include a legally binding norm prohibiting the design, development, production and use of fully AWSs (without meaningful human control);

Encouraging all states to open official negotiations within the United Nations (under the auspices of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons and/or within other appropriate United Nations conventions, bodies, agencies, and institutions) with the goal of developing regulations or other guidelines which will apply to the design, development, production and use of AWS.

Finding that this also enhances the need toward a global approach to AWSs.

***

Aware of:

- United Nations, Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS);[5]

- United Nations, Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001;[6]

- European Parliament Resolution of 12 September 2018 on autonomous weapon systems;[7]

[5] https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons__Informal_Meeting_of-_Experts_(2016) /ReportLAWS_2016_AdvancedVersion.pdf

[6] https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/#:~:text=It%20was%20adopted%20on%2010,or%20to%20affect%20civilians%20indiscriminately

[7] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_EN.html

- Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons, Communique of the Latin American and The Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons, 23 and 24 February, 2023;

- International Committee of the Red Cross Report, Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons, Geneva, September 2016;[8]

- International Committee of the Red Cross, Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 13-17 April 2015, Geneva;[9]

- XXth AIDP Congress, Criminal Justice and Corporate Business, Rome, 13-16 November 2019.[10]

***

The participants of the International Colloquium of the Section IV, held in Opatija, Croatia (7-8 December 2023), have adopted what follows:
Resolutions

1) In order to prevent and reduce AWS-related harms, it is necessary for international, regional, national legislators and other competent authorities to fully define AWS, as well as to develop regulations governing design, development, production and use of AWS.

2) Use of AWS must be regulated in advance by a law or a norm of equivalent binding force.[11]

3) States are urged to be more transparent about their use of AWS.

[8] ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, ICRC, Geneva, September 2016, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems

[9] ICRC (2015) *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS),* 13-17 April 2015, Geneva, https://www.icrc.org/en/document/lethal- autonomous-weapons-systems- LAWS

[10] https://www.penal.org/fr/aidp-xx-international-congress-penal-law-%E2%80%9Ccriminal-justice-and-corporate-business%E2%80%9D-rome-13-16-0.

[11] Article 36 of Additional Protocol I to the Geneva Conventions already requires member states to continue to assess whether the development or use of any new weapons would be prohibited by the Additional Protocol or by any other rule of international law. Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), Article 36, https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-additional-geneva-conventions-12-august- 1949-and

4) States must ensure that decisions taken by authorities to use AWS are based on transparent criteria that are subject to public scrutiny. States that use AWS must choose systems the functioning of which is fully transparent, explainable and traceable (white box AI).

5) Law or other equivalent norms related to the use of AWS must require that such systems have a high degree of technological reliability. A sufficiently precise regulation requiring appropriate evaluations and validations, independent of the AWS developer, must limit to the greatest possible extent the risk of bias or any form of discrimination in machine learning, coding errors, and other technological malfunction.

6) Laws or equivalent norms must require that before an AWS based on self-learning algorithms may be used in surveillance, selection of targets or any other purpose, the algorithms must be developed, trained, tested and used under human supervision (human-in- the-loop principle). These laws and equivalent norms must require a human evaluation before any action is taken to use an AWS.

7) States and law enforcement authorities must ensure that their personnel who operate AWS receive appropriate training in the proper use of the relevant AWS, as well as training with respect to the risk of error and bias. They must ensure that such personnel have a thorough knowledge of the dangers that AWS may pose to human rights.

8) Laws or equivalent norms must explicitly provide that where there is a human rights violation occurring as a result of any malfunction of an AWS, the company that designed or manufactured the system will incur criminal liability for fault or negligence[12] without excluding civil liability for defective products. Such laws or equivalent norms must also provide that investigations must be carried out to determine the cause of the violation.

9) Every state choosing to use AWS must have adequate penal laws in place, which must provide that individuals who use AWS in a manner inconsistent with relevant international and national legal standards will incur individual criminal responsibility before competent penal jurisdictions.[13]

10) States and regional and international human rights bodies must ensure that the use of AWS does not lead to serious human rights violations, *ius cogens* violations or other violations of relevant international law and international humanitarian law norms.

---

[12] AIDP XXI Congress: Artificial Intelligence and Criminal Justice, Section I: Traditional Criminal Law Categories and AI, Resolutions, approved by the Colloquium of Siracusa, 15-17 September 2022; AIDP XXI Congress: Artificial Intelligence and Criminal Justice, Section II: Penal Law and Criminalization in the Face of the Challenges of AI, Resolutions, approved by the Colloquium of Bucharest, 16-16 June 2023.

[13] AIDP XXI Congress: Artificial Intelligence and Criminal Justice, Section I: Traditional Criminal Law Categories and AI, Resolutions, approved by the International Colloquium of Siracusa, 15-17 September 2022.

11) States must ensure that they have fully implemented their international law-based legal obligations into their respective national legal orders, through laws, regulations, executive orders, military codes, or other types of binding domestic legal norms. Such domestic legislation must provide for different modes of liability, including, but not limited to, command responsibility.

12) States should establish independent research bodies to study, on a continuous basis, whether the use of AWS is being conducted in a manner consistent with all applicable international and national law norms and regulations.

13) It is necessary to identify and define specific modes for attributing criminal responsibility to the persons (both natural and legal persons) who are responsible for the development of AWS.[14] Such responsibility should not exclude persons, either natural or legal, who contribute to the causal chain of harm: from the designer, programmer, producer, seller, distributor to the end-users of the systems themselves.

14) States should strive to foster cooperation between policy and academic experts on the subject of international law and members of the military, law enforcement or intelligence forces who are tasked with developing and/or deploying AWS. Such cooperation should be conducted with the goal of ensuring that the use of AWS is always consistent with relevant international law norms, as well as with applicable domestic laws and regulations.

15) States must provide for extraterritorial application of their domestic laws in instances where AWSs are deployed or otherwise utilized abroad by states or by entities under state control.

16) States must develop appropriate conflict of law and conflict of jurisdictions laws which allow for the application of domestic law in instances where AWS are deployed or otherwise utilized abroad. Such laws may be developed at the domestic, regional and/or international level. Such laws must allow for the imposition of criminal responsibility (as well as civil liability) against those natural or legal persons within the relevant states control who are responsible for the misuse or malfunctioning of the AWS located outside the territory of the relevant state.

17) States must develop models of judicial cooperation in criminal matters, especially in order to collect evidence related to AWS-caused offenses. Such models of judicial cooperation may be particularly necessary for states which have not already concluded appropriate bilateral, regional, or international agreements on this subject.

[14] AIDP XXI Congress: Artificial Intelligence and Criminal Justice, Section I: Traditional Criminal Law Categories and AI, Resolutions, approved by the Colloquium of Siracusa, 15-17 September 2022.

18) States must ensure that they have extradition mechanisms in place in order to effectively prosecute and punish those responsible for AWS-caused offences.

XXIe CONGRÈS INTERNATIONAL DE DROIT PÉNAL
(Paris, 25-28 Juin 2024)

Thème: Intelligence Artificielle et Justice Pénale

1. Les categories traditionnelles de Droit Pénal et l'IA : crise ou palingénésie ?

2. Le Droit penal et l'incrimination face aux défis qu'engendre l'IA

3. L'IA et l'administration de la justice pénale : « police prédictive », « justice prédictive » et droit de la preuve

4. Perspectives internationales sur l'IA : les enjeux pour la coopération judiciaire, le droit international humanitaire et le droit pénal

Section I : Les categories traditionnelles de Droit Pénal et l'IA : crise ou palingénésie ?

Préambule

Considérant

- qu'avec des degrés d'autonomie différents, les systèmes d'intelligence artificielle (IA) soutiennent et remplacent de nombreuses activités humaines ;

- que les systèmes d'IA peuvent être réellement utiles à la société en général et aux autorités répressives en particulier, précisément aux fins des enquêtes sur les infractions pénales ;

- que les systèmes d'IA sont de plus en plus autonomes et peuvent être d'un fonctionnement imprévisible pour les personnes qui les conçoivent, qui les programment, qui les produisent, qui les distribuent et/ou qui les utilisent ;

Observant

- que les systèmes d'IA ont de très nombreux domaines d'application et que les faits illicites liés à leur mise en œuvre sont susceptibles de porter atteinte à divers intérêts, à des biens juridiques et à des droits fondamentaux ;

- que les systèmes d'IA pourraient en outre contribuer de plus en plus à la perpétration d'infractions pénales en servant d'« instruments » à cet effet, et qu'ils risquent de faciliter l'émergence de nouvelles infractions ;

Accordant une attention particulière :

- à la tendance croissante à déléguer totalement ou partiellement des décisions aux systèmes d'IA dans différents domaines d'activité, ce qui soulève la question de la responsabilité des personnes physiques ou morales pour les préjudices causés par le fonctionnement autonome de ces systèmes ;

- à l'autonomie des systèmes d'IA et à la possibilité, dont débattent les milieux universitaires, de considérer ces derniers comme des auteurs d'infractions ;

Gardant à l'esprit

- l'importance des réponses que le droit pénal est tenu d'apporter afin qu'il soit possible de prévenir et de punir les infractions commises par les systèmes d'IA, par leur intermédiaire, ou à leur encontre ;

- la gravité des préjudices et des risques liés aux applications de l'IA ;

- les principes fondamentaux qu'il est impératif de garantir en établissant et en appliquant les sanctions pénales (en celles compris les peines au sens large dont peuvent être passibles les personnes morales), par exemple le principe de légalité et le principe de culpabilité, qui est indissociable de la personnalité de la responsabilité pénale ;

- que les catégories traditionnelles de droit pénal et les modes de responsabilité pénale doivent être examinés et si nécessaire adaptés aux nouvelles exigences de protection ;

Tenant compte

- des « Lignes directrices en matière d'éthique pour une IA digne de confiance », que le Groupe d'experts de haut niveau sur l'intelligence artificielle a présentées à la Commission européenne le 8 avril 2019, et autres recommandations notables formulées par des organes internationaux (par exemple l'« Étude de faisabilité quant à un futur instrument du Conseil de l'Europe sur l'intelligence artificielle et le droit pénal » publiée le 4 septembre 2020 par le Comité européen pour les problèmes criminels, qui fait partie du Conseil de l'Europe) ;

- de la proposition de Règlement de l'UE sur l'intelligence artificielle (le « Règlement sur l'IA ») ; du travail du Comité du Conseil de l'Europe sur l'intelligence artificielle ; des activités des Nations Unies concernant l'intelligence artificielle ;

- des recommandations formulées lors du XIV^{e} Congrès international (Vienne, 1989) au sujet des problèmes juridiques et pratiques posés par la différence entre le droit criminel et le droit administratif pénal ; de celles qui ont été formulées lors du XVIII^{e} Congrès international (Istanbul, 2009) au sujet de l'élargissement des formes de préparation et de participation ; et de celles qui ont été formulées lors du XIX^{e} Congrès international (Rio de Janeiro, 2014), dont le thème était « Société de l'information et droit pénal » ;

Les participants au Colloque international de la Section I, qui s'est tenu à Syracuse, du 14 au 17 septembre 2022, ont adopté ce qui suit :

Recommandations

I. Sur la notion d'intelligence artificielle et l'octroi de la personnalité juridique aux systèmes d'IA, dont les degrés d'autonomie diffèrent

1. La notion d'IA recouvre une multitude de systèmes algorithmiques et robotiques interagissant avec l'environnement qui ont été mis au point à l'aide de plusieurs techniques (par exemple l'apprentissage automatique) pour atteindre des objectifs fixés par des humains. Il n'est par conséquent pas souhaitable de donner une définition générale de l'IA aux fins du droit pénal.

2. Toutefois, étant donné que les systèmes d'IA sont susceptibles de causer des dommages dans de nombreux secteurs (en lien par ex. avec les véhicules autonomes, les systèmes médicaux robotisés, ou encore les outils de trading ou de gestion logistique par IA), il conviendrait de tenir compte, pour que le droit pénal puisse assurer la protection des biens juridiques et des droits fondamentaux, non seulement des caractéristiques propres des différents systèmes d'IA, dont les degrés d'autonomie varient, mais encore des définitions juridiques émanant, ~~dans~~ chaque secteur concerné, de sources autres que le droit pénal.

3. En l'état actuel des choses, il n'existe aucune base ni aucune cohérence normative en matière de fonction des sanctions pénales pour pouvoir octroyer la personnalité juridique à des systèmes d'IA, lesquels ont des degrés d'autonomie variés.

4. D'une part, ontologiquement parlant, les agents humains se distinguent des systèmes d'IA : ces derniers n'ont en effet pas la conscience qui leur permettrait de choisir et d'évaluer des solutions à un problème ou un dilemme en tenant compte aussi du contexte dans lequel s'inscrivent des relations et opportunités d'ordre social ou éthique, avec la souplesse et la capacité nécessaires pour s'adapter à diverses situations et conditions, même casuelles ou nouvelles.

5. D'autre part, les peines dont seraient passibles de tels systèmes et agents technologiques ne satisferaient pas aux buts et fonctions des sanctions pénales et de leur application puisque la fonction d'intimidation serait vidée de tout sens par l'absence de conscience, chez ces systèmes et agents, de leur propre existence, passée, présente et future, mais aussi et surtout par leur absence de libre autodétermination, laquelle exclut également la fonction rétributive et la fonction de prévention, aussi bien spéciale que générale.

II. Sur la nécessité d'une réglementation, de normes et d'obligations extra-pénales

6. Pour prévenir et atténuer les préjudices liés à l'IA avant des réformes du droit pénal, ou tout au moins parallèlement à celles-ci, il faudrait que les législateurs et autorités compétentes, aux échelons tant international, régional que national, définissent dans leur intégralité les règles qui s'appliqueraient dans les nombreux secteurs dans lesquels des systèmes d'IA sont mis en œuvre (par exemple, comme indiqué plus haut, les secteurs dans lesquels des véhicules autonomes, des robots médicaux et chirurgicaux, des armes autonomes, etc. sont employés). Il faudrait élaborer une réglementation relative aux normes techniques, aux caractéristiques structurelles et aux conditions d'utilisation des systèmes d'IA et de leurs composantes.

7. Ces règles – qui doivent couvrir aussi bien les phases de la conception, de la production, de la distribution et de la vente que celle de l'utilisation elle-même des systèmes d'IA – devraient en outre prévoir des exigences concrètes quant à l'adaptation des systèmes en cas d'avertissements ou d'alertes, ce qui serait la condition sine qua non permettant d'invoquer ensuite le droit pénal pour sanctionner des dommages liés à l'IA.

8. Ces règles pourraient prévoir des injonctions – comme c'est déjà le cas dans des secteurs comportant une multitude de risques (par ex. la santé et la sécurité sur le lieu de travail, la protection de l'environnement) – dont la violation ou l'inobservation seraient passibles de sanction, dans le respect du principe de l'*ultima ratio*.

III. Sur la nécessité d'une protection pénale des biens juridiques

9. Il est nécessaire de reconnaître que l'intervention du droit pénal au sens large, en toute rationalité et proportionnalité, est indispensable pour empêcher et punir des faits d'atteinte ou de mise en danger que des systèmes d'IA pourraient causer à des intérêts, des biens juridiques et des droits fondamentaux ; en effet, selon les catégories traditionnelles du droit pénal, de tels faits seraient constitutifs d'une infraction pénale s'ils étaient commis par des personnes physiques ou morales. Il s'ensuit que tels faits ne peuvent pas rester impunis simplement pour avoir été commis par les systèmes susmentionnés, par leur intermédiaire ou à leur encontre.

10. Il est nécessaire de retenir et de définir des modes précis d'attribution de la responsabilité aux personnes (aussi bien physiques que morales) qui sont « derrière » les systèmes d'IA (c.-à-d. les acteurs des différentes phases du cycle de vie des systèmes d'IA : concepteurs, fournisseurs, importateurs, distributeurs, utilisateurs, etc.), à commencer par les propriétaires de ces systèmes et par les personnes qui décident quelle utilisation concrète en sera faite, décision qu'elles prendront en fonction de leur intérêt et à leur profit : ces personnes doivent donc être tenues juridiquement responsables et sanctionnées, et pas seulement sous l'angle du droit pénal.

11. La responsabilité des personnes décrites au point précédent n'exclut en rien celle des autres personnes (physiques ou morales) ayant apporté une contribution dans la

chaîne de causalité du préjudice, qu'il s'agisse des concepteurs, programmeurs, producteurs, revendeurs ou distributeurs des systèmes, voire des utilisateurs finaux eux-mêmes.

12. Il faut opérer une distinction entre :

a. Les systèmes d'IA utilisés dans le cadre d'activités illégales. Il s'agira ici essentiellement de comportements malveillants, lesquels posent moins de problèmes d'attribution de la responsabilité pénale car théoriquement les systèmes d'IA ne diffèrent pas d'autres instruments et moyens de commission d'une infraction.

Il faut toutefois examiner deux situations :

a.1. lorsque l'utilisation du système aux fins de l'activité illicite prévue n'a pas donné les résultats escomptés, deux principes classiques doivent s'appliquer : *aberratio ictus* et *aberration delicti*. Le simple fait que l'objet lésé n'ait pas été celui qui était prévu est inopérant dans la mesure où ça n'a aucune incidence sur les éléments constitutifs de l'infraction pénale (par ex. tuer une personne au lieu d'une autre n'a pas d'incidence sur la réalisation du crime de meurtre lorsque l'agent a agi avec l'intention de commettre un meurtre). Il est alors préférable de fonder la responsabilité pénale pour une infraction différente de celle qui était projetée sur la possibilité de prévoir concrètement que l'action lancée par le système d'IA pouvait avoir des résultats différents : ce sont les principes de la responsabilité pour négligence (voir paragraphe 4 plus bas) qu'il faut retenir ;

a.2. les systèmes d'IA pouvant être utilisés pour des comportements particulièrement nuisibles ou dangereux, ils sont susceptibles de renforcer et d'aggraver les dommages causés (comme ça se produit avec les TIC). Les conséquences pouvant être en effet très éloignées des actes nuisibles qui les ont causées, il est d'autant plus difficile d'intervenir *post-factum* pour empêcher ou à tout le moins faire cesser ou atténuer ceux-ci. Il faudrait par conséquent envisager d'incriminer, en tant qu'infractions préparatoires autonomes, le fait de concevoir, programmer, produire, distribuer, vendre et acquérir, illégalement, des algorithmes, logiciels et systèmes d'IA « malveillants ». Il conviendrait que ces mesures pénales ne s'appliquent qu'aux systèmes d'IA à hauts risques pour certains biens juridiques de premier plan (par exemple la vie, l'intégrité physique ou la liberté d'autrui), et ce seulement en cas de danger précis, réel et actuel (à propos des conditions requises pour incriminer les actes préparatoires, voir la résolution de la Section I du XVIII^{e} Congrès de l'AIDP, tenu à Istanbul en 2009).

b. Les systèmes d'IA utilisés dans le cadre d'activités légales. Se posent ici les questions les plus épineuses en matière de « risque acceptable », lequel devrait s'inscrire dans le cadre d'obligations spécifiques en matière de sécurité et de mesures de précaution applicables aux activités de conception, mise au point, production, distribution et vente ainsi qu'à l'utilisation des systèmes d'IA, la mise en place d'un tel cadre étant souhaitée. Dans ce domaine, il faut adapter les différents modes de responsabilité pénale pour trouver un équilibre entre les diverses formes de responsabilité pour négligence et les

caractéristiques techniques des systèmes d'IA, à savoir : 1) leur autonomie, 2) l'imprévisibilité concrète de leurs décisions et de leur fonctionnement ; 3) l'opacité de leurs mécanismes de régulation, 4) la complexité de leurs processus de programmation, mise au point, production, mise à jour et maintenance.

IV. Sur l'adaptation des modes d'attribution de la responsabilité aux caractéristiques des systèmes d'IA, plus précisément à leur degré d'autonomie

13. Il faut tout d'abord opérer une distinction – en fonction des paliers d'automatisation et d'autonomie établis pour les applications d'IA dans plusieurs domaines – entre les différents niveaux d'autonomie décisionnelle et opérationnelle dont les systèmes d'IA disposent : cela va des systèmes dans lesquels le mode de fonctionnement « automatique » laisse aux agents humains une marge de contrôle importante sur eux, aux systèmes qui sont véritablement « autonomes », dans le fonctionnement desquels les agents humains ne peuvent intervenir qu'à distance, aussi bien dans le temps que dans l'espace, ces systèmes-là « prenant des décisions » sur la base des informations collectées et d'algorithmes qui s'adaptent au fur et à mesure du fonctionnement : il y a donc une marge structurelle d'imprévisibilité des résultats concrets.

14. S'agissant des différents types de systèmes d'IA, il est absolument indispensable que des règles et normes spécifiques de fonctionnement soient élaborées, comme le laisse entendre la proposition de Règlement de l'UE sur l'intelligence artificielle (voir paragraphe 2).

15. Le plus urgent, c'est d'adapter les catégories traditionnelles du droit pénal en ce qui concerne les systèmes d'IA ayant le plus haut degré d'autonomie : la tendance actuelle en matière d'innovation et d'expérimentation technologique leur étant favorable dans de nombreux secteurs, leur rôle se renforcera indéniablement dans un avenir proche.

16. Dans cette perspective, le domaine de la responsabilité pénale des personnes morales peut être une précieuse source d'information, tout comme éventuellement les domaines de la responsabilité du fait des produits défectueux et de la responsabilité en matière de protection de la santé et de la sécurité sur le lieu de travail.

17. Les règles juridiques qui sont en vigueur dans ces domaines et souvent harmonisées à l'échelon européen sont assorties de principes susceptibles d'être étendus, moyennant les adaptations qui s'imposent, à la réglementation des infractions pénales liées à l'IA. Dans ces domaines où il existe déjà des réglementations, il est imposé de procéder en amont à une évaluation des risques inhérents aux activités spécifiques exécutées, avec une marge de risque acceptable et des obligations connexes en matière de prévention et d'endiguement des risques (voir paragraphe 2), tout particulièrement à l'égard des sources de mise en danger et de préjudices.

18. Un devoir d'action, tout particulièrement en cas de déclenchement d'alertes, est imposé à diverses catégories de personnes (humaines) en fonction de leurs compétences respectives, par exemple aux utilisateurs et aux personnes occupant une position de

garant. Ces personnes doivent immédiatement adapter les règles et mesures de sécurité de leur activité, voire, si nécessaire, interrompre celle-ci.

19. Sur la base de ces principes établis, les recommandations ci-après peuvent être formulées quant à la structure de la responsabilité pénale pour les dommages liés à l'IA :

i. Responsabilité pénale des personnes physiques. Cette forme de responsabilité impose de désigner les personnes occupant la position de garant à l'égard des compétences et fonctions confiées aux systèmes d'IA. Il faut commencer par examiner l'ensemble des acteurs participant aux différentes phases d'utilisation des systèmes d'IA, en ce compris les utilisateurs finaux. Il faut ensuite analyser le rôle des différents gestionnaires (cadres moyens et supérieurs) ainsi que celui des responsables de la conformité dans les organisations de grande taille, et examiner quelles sont les obligations positives officielles de chacune de ces personnes, que ce soit en termes techniques, organisationnels ou de supervision.

La responsabilité pénale pour négligence doit respecter les principes généraux du droit pénal, en l'occurrence le principe de culpabilité personnelle : en effet la relation objective entre l'imputation causale à l'agent humain et la commission de l'infraction par le système d'IA est insuffisante puisque la prévisibilité et l'évitabilité de l'acte illicite sont également requises. La responsabilité pénale pour négligence – ne pas avoir agi différemment compte tenu de ce qu'il était possible de faire – ne relève pas tant des faits ou actes spécifiques et concrets qui se sont produits que du régime de la « faute de l'organisation », c'est-à-dire de la façon dont l'agent artificiel est structuré et dont il fonctionne. Pour évaluer les risques qu'engendrent les activités des systèmes d'IA, il faut tenir compte aussi des informations concernant les résultats concrètement « imprévisibles » au cas par cas, base sur laquelle se fonde l'obligation de préparation raisonnable et d'adoption de mesures de surveillance et d'endiguement actualisées en permanence, obligation qui est imposée à la personne physique responsable, qui reste tenue de rendre des comptes (transparence) en tant que propriétaire ou dirigeante de l'entité utilisant le système d'IA dans son intérêt et à son profit.

ii. Responsabilité pénale des personnes morales. Étant donné que les systèmes d'IA sont en grande partie produits ou utilisés par des personnes morales, il est nécessaire que ces dernières soient tenues responsables des infractions commises par lesdits systèmes, par leur intermédiaire ou à leur encontre.

À cet égard, à supposer que des normes publiques précises de conduite et de conformité soient adoptées (voir paragraphe 2), les peines dont les personnes morales seraient passibles, lesquelles devraient être proportionnelles aux infractions commises par les systèmes d'IA, par leur intermédiaire ou à leur encontre, et au degré de négligence de l'entité, pourraient relever d'un mode de responsabilité fondé sur la faute de l'organisation. Un tel mode de responsabilité amène à imputer de manière subjective la responsabilité, les faits incriminés, à la personne morale en cas d'infractions causées par l'absence, l'in-

suffisance ou l'imperfection des mesures structurelles et préventives à exécuter et actualiser à la lumière de l'évaluation des risques découlant expressément des activités confiées aux systèmes d'IA et au demeurant accomplies par eux, dans l'intérêt de ladite personne morale ou à son profit. Il convient d'éviter les modes de responsabilité de plein droit.

Il faudrait œuvrer en faveur de l'adoption d'un nouveau mode de responsabilité pénale autonome des personnes morales, qui ne soit pas basé sur la responsabilité individuelle des personnes physiques, et ce car une personne morale peut être tenue responsable même si la personne physique qui a causé le dommage ne peut être sanctionnée à titre individuel en raison de conditions ou circonstances particulières ou bien car elle ne peut pas être expressément identifiée. Il est en effet suffisant d'établir qu'un acte objectivement et généralement illicite a été commis dans l'intérêt de l'organisation ou à son profit.

Dans les systèmes juridiques nationaux où la responsabilité des personnes morales se fonde sur une liste fermée d'infractions, il est recommandé d'élargir cette liste à des infractions pénales susceptibles d'être commises par des systèmes d'IA, par leur intermédiaire ou à leur encontre.

*V. Sur les mesures préventives et les sanctions punitives applicables aux personnes physiques et aux personnes morales qui sont « derrière » les systèmes d'intelligence artificiel***le**

20. Les sanctions dont sont passibles les personnes physiques dans les divers systèmes juridiques, en ce compris l'emprisonnement, et les personnes morales, pour lesquelles il pourra s'agir de sanctions à caractère administratif, du reste infligées à des fins punitives, en ce compris des amendes et la suspension des activités par le truchement desquelles l'infraction a été commise, devraient coïncider avec celles qui sont infligées pour le type d'infraction commise et respecter les principes ayant cours dans chaque système juridique, à savoir la proportionnalité et la personnalité de la sanction. S'agissant des personnes morales, ces sanctions, en sus des mesures pécuniaires, devraient être assorties de l'injonction de corriger le système de contrôle de la conformité et de contrôle interne de la personne morale, ainsi que de la possibilité d'ordonner une période de contrôle public de la personne morale afin de vérifier que les normes imposées sont respectées.

21. Compte tenu de la gravité des préjudices liés à l'IA, il est recommandé que des mesures préventives soient adoptées pour pouvoir empêcher ou atténuer efficacement les conséquences préjudiciables liées aux systèmes d'IA (entre autres saisies, confiscation, contrôle judiciaire, mesures d'interdiction).

22. Il convient de mettre en exergue le rôle important des peines non pécuniaires, par exemple l'interdiction d'exercer des activités précises et la confiscation. Cette dernière permet de prendre des mesures directes à l'encontre du système d'IA avec lequel ou par l'intermédiaire duquel l'infraction a été commise, et ce sans qu'il faille pour autant

lui octroyer le statut de personne morale ou la responsabilité pénale (voir plus haut, paragraphe 1).

VI. Systèmes de garantie complémentaires

23. Compte tenu des problèmes et difficultés qui ne manqueront pas d'accompagner la mise en œuvre d'un système effectif de responsabilité pénale à l'égard des personnes physiques et des personnes morales qui sont « derrière » les systèmes d'IA pour des infractions commises par ces derniers, par leur intermédiaire ou à leur encontre, il est recommandé de mettre en place des systèmes de garantie complémentaires.

24. Il pourrait s'agir d'autorisations et d'homologations administratives mais aussi de moyens de recours au civil.

24. Il pourrait aussi s'agir, en lieu et place des poursuites et/ou sanctions pénales, de modes de conformité, d'interventions de justice restaurative et d'accords avec les victimes et les pouvoirs publics compétents.

Section II : Le Droit penal et l'incrimination face aux défis qu'engendre l'IA

Préambule

Considérant

- que l'intelligence artificielle, l'une des dernières avancées issues de la révolution numérique, a déjà atteint un stade de développement remarquable en cette troisième décennie du 21e siècle et est déjà largement utilisée dans de nombreux secteurs de la société ;

- que bien qu'aucune définition de cette technologie ne fasse l'unanimité, l'on s'accorde implicitement à dire qu'il s'agit d'une multitude de systèmes informatisés qui, par l'accumulation, le traitement et l'analyse de données dans leur contexte, sont capables d'agir de façon autonome et/ou d'aider à la prise de décisions visant à atteindre des objectifs précis ;

- que le potentiel de transformation de ces technologies a des incidences dans de nombreux domaines et sphères sociales, suscitant dans son sillage d'importants avantages et opportunités ;

- que l'intelligence artificielle fait par ailleurs peser des risques sur des intérêts individuels et collectifs et porte atteinte à ces intérêts ;

Observant

- que de nombreuses décisions habituellement adoptées par des humains en connaissance de cause commencent à être automatisées grâce à l'utilisation de ces technologies, ce qui a des incidences sur différents domaines et différents intérêts ;

- qu'en ce qui concerne par exemple les véhicules autonomes, les services de santé, les marchés financiers, les médias et d'autres secteurs encore, l'utilisation de cette technologie est irrépressible et un avenir sans elle semble impossible ;

- que la promesse d'une IA à la fois efficace et objective engendre l'essor de ces technologies sans aucune évaluation de leur véritable nécessité ni des risques qu'elles sont susceptibles de susciter ;

Gardant à l'esprit

- que les progrès récemment accomplis en ce qui concerne les grands modèles de langage et d'autres systèmes d'IA, par exemple l'apprentissage automatique et l'apprentissage en profondeur, ont mis en lumière la nécessité d'une réglementation, notamment de protocoles de sécurité, pour contrôler l'évolution de ces technologies en raison de leurs effets et des risques qu'elles posent ;

- que la mise au point des systèmes d'IA, en particulier l'entraînement de leurs algorithmes, fait appel à l'accumulation et à l'utilisation de données et de grandes quantités d'informations, ce qui représente un risque dont il faut absolument tenir compte ;

Soulignant

- l'inquiétude croissante que suscite le risque que des utilisations malveillantes ou imprudentes de l'IA causent des dommages dans des domaines dans lesquels l'IA commence à être déjà fortement présente ;

- que nombreux sont les pays où l'utilisation de systèmes d'IA a nui à d'importants intérêts tels que la vie, la santé et la vie privée ;

Reconnaissant

- que l'émergence de nouvelles infractions et de nouveaux intérêts dignes de protection en droit pénal amènera les États à adapter les lois pénales en ce qui concerne l'IA ;

- qu'il faut étudier si les États et la communauté internationale apportent une réponse juridique adéquate aux défis que pose l'IA ou s'il faut procéder à des réformes et adaptations, soit en modifiant expressément certaines dispositions soit en créant de nouvelles formes d'incrimination ;

Tenant compte

- de la Recommandation du Conseil sur l'intelligence artificielle, adoptée par l'OCDE le 22 mai 2019 ; de la proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes législatifs de l'Union (législation sur l'intelligence artificielle), qui a été présentée par la Commission européenne le 21 avril 2021 ; de la résolution du Parlement européen du 6 octobre 2021 sur l'intelligence artificielle en droit pénal et son utilisation par les autorités policières et judiciaires dans les affaires pénales (2020/2016(INI)) ; des « Lignes directrices en matière d'éthique pour une IA digne de confiance », que le Groupe d'experts de haut niveau sur l'intelligence artificielle a présentées à la Commission européenne le 8 avril 2019 ; et de la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme (adoptée par le Comité des Ministres le 8 avril 2020, lors de la 1373e réunion des Délégués des Ministres) ;

- des débats et résolutions de précédentes éditions du Congrès international de droit pénal, en particulier les résolutions du XIVe Congrès, sur le droit pénal et les techniques biomédicales modernes, ainsi que sur les problèmes juridiques et pratiques posés par la différence entre le droit criminel et le droit administratif pénal ; les résolutions du XIXe Congrès international de droit pénal, sur le thème « société de l'information et droit pénal » ; et les projets de résolutions du XXIe Congrès international de droit pénal sur la Section I (Droit pénal – Partie générale) et sur la Section III (l'IA et l'administration de la justice pénale) ;

Les participants au Colloque international de la Section II, qui s'est tenu à Bucarest, du 14 au 16 juin 2023, ont adopté ce qui suit :

Résolutions

I. Réglementation de l'IA et contrôle de l'application des règles

1. Dans le sillage de l'essor et de l'expansion de l'IA, l'atteinte à des intérêts individuels et collectifs ainsi que les risques pesant sur ces intérêts ont augmenté et continueront vraisemblablement de le faire. Il est donc impératif d'établir des règles adéquates relatives à la production, à la mise en œuvre et à l'utilisation des systèmes d'IA. Il faut, pour répondre à ces défis, tenir compte des différents contextes concernés et de tous les outils réglementaires en vigueur, publics ou privés, ainsi que de la nature et de la fonction de chacun.

2. Les répercussions de l'IA étant mondiales, la réponse apportée doit être internationale pour que les intérêts individuels et collectifs qui sont en jeu puissent être dûment protégés. Les États devraient tenir compte des normes internationales dans leur réglementation nationale et dans l'exécution de celle-ci.

3. Comme c'est arrivé avec d'autres innovations technologiques et socio-économiques, par exemple internet ou les nouvelles avancées en matière de modification génétique et de neurosciences, l'irruption de l'IA impose de revoir les aspects généraux du système de justice pénale et en particulier la liste des infractions existantes en droit pénal pour s'assurer qu'avec les règles dont ils disposent, les États pourront surmonter les difficultés liées à l'utilisation de cette technologie.

4. Il ne faut pas séparer les débats sur la transformation du droit pénal sous l'influence de l'IA et sur le rôle que le système de droit pénal pourrait jouer en lien avec ces nouvelles technologies, des débats éthiques tenus à l'échelon international au sujet de l'essor de l'IA, pas plus que des débats que d'autres branches du droit tiennent au sujet de la réglementation de l'IA. Les législateurs devront se replacer dans ce contexte global lorsqu'ils réfléchiront au rôle précis que joue le droit pénal pour éviter les préjudices causés par l'utilisation de l'IA.

5. Les législateurs devront réformer les infractions existantes dans les cas où l'IA modifie le risque pesant sur des intérêts existants ou crée de nouveaux modes de commission d'infractions non encore couverts dans la législation en vigueur. Il faudra néanmoins créer de nouvelles infractions lorsque verront le jour, sous l'impulsion de l'essor de cette technologie, de nouveaux intérêts individuels et collectifs dignes de protection mais non encore couverts par la législation en vigueur.

II. Incrimination et protection des intérêts liés à l'IA

6. L'essor de l'IA pourrait faire émerger de nouveaux intérêts dignes de protection. Il se pourrait en outre que les systèmes d'IA agissent sur la portée et la valeur d'intérêts qui ne sont à l'heure actuelle pas considérés comme étant dignes de protection par le droit pénal. Lorsque les lois pénales ne permettent pas de dûment protéger lesdits intérêts, il faut que de nouvelles incriminations soient créées pour sanctionner en toute proportionnalité des comportements portant atteinte à ces intérêts. Ce doit être le cas uniquement lorsqu'il n'est pas possible de dûment protéger ces intérêts avec des moyens plus doux que le droit pénal.

7. Lorsque la transformation de l'IA entraîne l'émergence de nouveaux intérêts en substance similaires à d'autres intérêts depuis toujours considérés comme étant dignes de protection, il n'est pas nécessaire de créer de nouvelles sanctions pénales. Il est en revanche préférable d'adapter l'interprétation des infractions existantes, pour autant que le respect strict du principe de légalité le permette.

8. Les législateurs devront décider si l'essor de cette technologie impose de mettre en place une protection pénale expressément applicable à des intérêts individuels ou collectifs liés à la technologie de l'IA elle-même. Bien qu'il soit encore trop tôt pour se prononcer, il se pourrait que ce soit le cas des données sur lesquelles reposent les algorithmes, parfois de la fonctionnalité des systèmes d'IA eux-mêmes, des intérêts collectifs liés à la sécurité et à la fiabilité de leur conception et de leur application, voire des intérêts liés aux robots.

9. Certains systèmes d'IA, par exemple ceux qui sont utilisés dans les infrastructures critiques, sont essentiels pour garantir des intérêts déjà protégés. Dans la mesure où la Convention de Budapest prévoit l'incrimination d'actes portant atteinte aux systèmes informatiques et dans la mesure où l'IA peut être considérée comme telle, il pourrait ne pas être nécessaire d'adopter de nouvelles infractions. Afin de ne pas soulever des doutes d'interprétation, il pourrait être souhaitable de modifier certaines infractions pénales afin d'ajouter les systèmes d'IA parmi les différents types de systèmes informatiques.

10. S'il n'est pas prouvé que divers autres intérêts sont en jeu, et compte tenu du niveau actuel de développement de ces technologies, cela signifiera alors que l'IA et les systèmes robotiques n'ont pas besoin d'une protection différente – eu égard à leur valeur économique ou fonctionnelle – de celle dont bénéficient d'autres systèmes informatiques.

III. Motifs de légitimation et techniques d'incrimination

11. Le droit pénal ne doit pas jouer un rôle de premier plan dans la réglementation de l'IA. Comme de par sa nature le droit pénal est un instrument de répression particulièrement coercitif, il ne doit intervenir qu'en dernier recours et son usage doit être limité à la répression des actes les plus graves et les plus préjudiciables.

12. Les législateurs ne doivent pas adopter de nouvelles infractions au seul motif qu'une IA a été utilisée. De nombreuses infractions pénales peuvent avoir comme mode de commission du comportement incriminé des systèmes d'IA. C'est uniquement s'il s'avère que les actes commis avec des systèmes d'IA ont une portée différente en termes de nuisibilité ou de risque qu'il sera nécessaire d'adopter de nouvelles infractions pénales.

13. L'automatisation grâce à l'IA des processus décisionnels axés sur les données déplace le moment déterminant de l'intervention humaine vers les phases – très éloignées du préjudice – de la conception et de la mise en œuvre des algorithmes. Il conviendrait par conséquent que la responsabilité des personnes physiques et des personnes morales qui y sont associées se concentre sur ces phases-là. Il faut à cet effet prendre en considération les obligations légales qui ont été établies dans d'autres branches du droit.

14. Les systèmes de justice pénale sont conçus pour avoir un effet dissuasif sur les délinquants potentiels, c'est-à-dire pour les empêcher de commettre des infractions. Si le moment déterminant, en termes de risques liés à l'IA, est celui de la conception et de la mise en œuvre, il faut alors envisager d'adopter des infractions pour empêcher la commission d'actes illégaux à ces moments-là. Cela peut être fait à l'aide de mesures de protection anticipative, par l'adoption d'infractions de mise en danger permettant de punir le non-respect de certaines obligations liées à des intérêts spécifiques dignes de protection. Par ailleurs, à l'instar des dispositions en vigueur en matière de responsabilité pénale des personnes morales, des obligations réglementaires pourraient être établies en ce qui concerne expressément la conception et la mise en œuvre des systèmes d'IA, obligations dont la violation pourrait engager la responsabilité pénale de son auteur.

15. Il faut adopter des infractions de mise en danger lorsque de par leur conception et leur mise en œuvre, des systèmes d'IA à risque font peser une menace importante sur des intérêts protégés. En outre, les conséquences juridiques liées à ces infractions doivent être proportionnelles au niveau de risque pesant sur les intérêts en jeu. À cause de la complexité de conception des systèmes d'IA et des différentes approches de la réglementation de ces outils, il ne faut pas adopter d'infractions de mise en danger avant d'avoir fait le point, pour chaque système juridique, sur les mesures d'autorégulation ou les règles administratives en matière de contrôle et de sécurité de l'IA. Ces règles doivent servir à repérer les actes à risque susceptibles d'être passibles de poursuites pénales.

16. De nouvelles infractions pourraient être adoptées pour punir le recours abusif à des systèmes d'IA légaux existants et la transformation de ces derniers lorsque du fait de la modification de la conception ou de l'objet de l'IA, de nouveaux risques émergent.

17. Dans les systèmes juridiques où seuls les cas de commission par omission expressément prévus sont punis (*numerus clausus*), peut-être faudrait-il réformer les lois pénales. En raison de la complexité de conception de l'IA et de l'implication d'une multitude de parties tout au long du cycle de vie de l'IA, il est dans la plupart des cas extrêmement difficile de prouver que quiconque avait connaissance dès la conception qu'un préjudice allait se produire. Des infractions par négligence qui sont fondées sur la violation du devoir de vigilance pourraient être adoptées si elles s'avéraient nécessaires aux fins de la protection des intérêts concernés.

18. Étant donné que les systèmes d'IA sont dynamiques et que leurs performances dépendent de l'incorporation ou de la collecte de données modifiant leurs résultats, il se peut que les processus de gestion des risques mis en place dans d'autres branches du système juridique puissent s'appliquer tout au long du cycle de vie de l'IA. Partant, les lois pénales pourraient incriminer, si nécessaire, la violation des règles relatives au défaut de surveillance et de contrôle des systèmes d'IA et de telles obligations pourraient s'appliquer à différents sujets tout au long du cycle de vie de l'IA.

19. La technologie de l'IA et ses applications étant évolutives, peut-être faudrait-il que les systèmes de justice pénale puissent s'adapter afin qu'il soit possible d'ajuster la proportionnalité des peines à la sévérité des préjudices que l'IA est susceptible de causer. Néanmoins, les législateurs ne doivent pas adopter de circonstances aggravantes uniquement au motif qu'une infraction a été commise à l'aide de l'IA. L'adoption de nouvelles formes d'aggravation des peines ne pourra être envisagée que si les circonstances aggravantes en vigueur ne permettent pas de dûment répondre à la gravité des dommages causés ou à l'importance des intérêts touchés du fait de l'utilisation de l'IA. Le principe de proportionnalité devra être respecté en toutes circonstances.

IV. Incrimination et protection d'intérêts spécifiques face aux risques engendrés par l'IA

20. Étant donné que les lois pénales n'énoncent habituellement pas de modes précis de commission d'infractions contre la vie et la santé, il ne semble pas nécessaire de réformer ces infractions pour pouvoir protéger ces intérêts dans les cas où le mode de commission utilisé est un système d'IA, pas plus qu'il ne semble nécessaire de modifier le système de gradation des infractions. Il n'en demeure pas moins que dans des domaines spécifiques, par exemple la conduite autonome, le droit pénal devra être attentif à l'évolution du cadre réglementaire sur la base duquel le « risque acceptable » sera déterminé et, partant, la responsabilité.

21. Si la conduite autonome devient chose courante, les infractions à la sécurité routière pourraient subir des modifications non négligeables et de nouvelles infractions être adoptées en lien avec l'apparition de nouveaux comportements présentant un risque pour la vie et la sécurité routière, outre celles qui existent déjà pour la conduite humaine.

22. Tout comme la révolution en matière de modification génétique a entraîné l'apparition de nouvelles infractions sanctionnant les modifications génétiques capables de causer une destruction massive, l'évolution technologique peut imposer d'incriminer, dans un futur proche, la création, la mise au point et l'utilisation d'outils d'IA à haute capacité de destruction, par exemple certaines armes autonomes, des drones ou des robots aptes à causer d'énormes dommages, en particulier s'ils ne sont plus soumis à un contrôle humain.

23. Les systèmes d'IA recueillent d'immenses quantités d'informations sur lesquelles ils s'appuient pour accomplir les tâches qui leur sont confiées, ce qui fait peser de nouvelles menaces sur des intérêts classiques. Cet essor de la technologie de l'IA impose de revoir les infractions liées au droit à la vie privée et à d'autres intérêts personnels, et de réexaminer l'interprétation du droit à la vie privée comme étant un bien seulement individuel alors qu'il conviendrait de prendre en compte sa dimension collective.

24. L'incrimination d'actes consistant à recueillir illégalement des données à caractère personnel ne devrait pas être uniquement fondée sur la protection d'intérêts tels que la vie privée. L'utilisation de l'IA dans le cyberespace pourrait ouvrir la porte à la collecte

massive de données en vue de la commission de cybercrimes susceptibles de porter atteinte à des intérêts tels que le droit de propriété. Les États devraient réfléchir à la question de savoir s'il faut adopter des infractions pénales pour sanctionner la collecte massive et illégale de données et des actes préparatoires similaires en vue de la commission d'infractions graves lorsque ceux-ci font peser des risques concrets sur ces intérêts et seulement en l'absence d'outil juridique moins coercitif.

25. Le fait que des images et des données à caractère personnel soient accessibles dans le cyberespace alors même que l'IA générative est capable de transformer des images, des vidéos et des fichiers audio, représente un danger pour des intérêts tels que la réputation, l'honneur ou la liberté sexuelle. Il est nécessaire d'examiner si les lois pénales en vigueur permettent de punir des comportements nuisibles pour la dignité humaine, la réputation et la liberté sexuelle : par exemple la diffusion d'hyper-trucages, notamment à caractère sexuel, ou de pédopornographie.

26. Les IA génératives, par exemple les grands modèles de langage et des outils similaires, peuvent faciliter la tromperie, les menaces et la coercition susceptibles de venir porter atteinte à différentes phases de la formation de la volonté, et de mettre ainsi en danger des intérêts dignes de protection. Il ne semble toutefois pas pertinent de créer de nouvelles infractions pénales puisque les lois pénales en vigueur couvrent déjà les actes les plus préjudiciables : il conviendrait donc d'avoir recours à d'autres moyens de contrôle de ce type d'outils pour empêcher les comportements moins risqués.

27. La popularisation des algorithmes de gestion des risques dans des domaines tels que la santé, le recrutement, la justice, l'octroi de crédits et prêts et bien d'autres encore, a mis en évidence l'existence de biais discriminatoires dans certaines décisions prises par les systèmes d'IA. Outre les infractions pénales qui permettent déjà de sanctionner des décisions particulièrement discriminatoires, d'autres branches du système juridique, notamment le droit civil et le droit administratif, sont plus à même de combattre le problème de la discrimination algorithmique.

28. L'utilisation de l'IA dans le cyberespace peut faciliter et renforcer l'existence d'atteintes à la propriété et à d'autres intérêts dignes de protection. Toutefois, compte tenu des règles en vigueur en matière de cyberfraude et autres cyberinfractions à la propriété, il ne sera pas nécessaire, du moins à court terme, d'adapter ces infractions pour couvrir celles qui sont perpétrées à l'aide des systèmes d'IA.

29. Avec certaines lois pénales punissant la production, la vente et l'obtention pour utilisation, l'importation, la diffusion ou autres formes de mise à disposition de dispositifs principalement conçus ou adaptés pour permettre la commission de l'une des infractions établies conformément aux articles 2 à 5 de la Convention de Budapest, il est d'ores et déjà possible de sanctionner la création, la mise au point et la vente de systèmes d'IA

conçus ou adaptés pour permettre la commission desdites infractions pénales. Par conséquent, si l'on considère que l'IA est un dispositif, y compris un programme informatique, au sens de l'article 6 de la Convention de Budapest, il n'est pas nécessaire d'adopter de nouvelles infractions en prévision d'éventuelles poursuites pénales.

30. Dans la sphère socio-économique et financière, on signale déjà une prolifération des systèmes algorithmiques de prise de décisions et du recours à l'IA pour le trading. Le risque que l'utilisation malveillante ou imprudente des systèmes d'IA puisse avoir un grave impact sur les marchés est évident, mais les intérêts en jeu pourraient être mieux protégés par des mesures réglementaires préventives d'ordre économique, administratif et commercial plutôt que par l'adoption d'infractions pénales autres que celles qui existent déjà pour punir les délits d'initiés et autres comportements similaires. Lorsque les systèmes d'IA sont utilisés pour manipuler les marchés, alors le droit pénal devrait être révisé pour pouvoir apporter une réponse proportionnée.

31. Les préoccupations quant à l'impact de la désinformation, au lendemain tout d'abord de certaines campagnes électorales puis de « l'infodémie » qui a eu lieu durant la crise de la Covid-19, ont amené bien des États à créer des infractions pénales pour sanctionner ces comportements. Le fait que l'IA puisse amplifier l'impact de la désinformation, soit par l'automatisation de sa diffusion ou par le recours à des technologies de manipulation des vidéos, des images, du son et des textes peut renforcer cette tendance. Incriminer la désinformation ne se justifiera que si c'est aux fins de la protection des droits fondamentaux des sociétés démocratiques et que si cette incrimination ne met pas en péril la liberté d'expression.

Section III : L'IA et l'administration de la justice pénale : « police prédictive », « justice prédictive » et droit de la preuve

1.1 Préambule

Conscients que dans la société contemporaine l'intelligence artificielle (IA) se développe rapidement en divers endroits du monde. Déjà omniprésente dans la vie des citoyens de certains pays, elle est peut-être en passe de s'ancrer dans le quotidien d'une grande partie de la population mondiale.

Notant qu'en tant qu'innovation technologique, l'IA pousse les consommateurs à acquérir de nouveaux produits, contribuant ainsi à la croissance économique mondiale. Elle joue de ce fait un rôle non négligeable dans le soutien, voire l'expansion de l'économie de marché libérale et du système économique capitaliste.

Notant que les entreprises qui créent et commercialisent des systèmes d'IA ont fréquemment leur siège dans des pays développés de l'hémisphère nord et tentent souvent de s'implanter sur des marchés partout dans le monde.

Notant que la « fracture numérique » accentue les inégalités sociales au sein de la population. La « fracture de l'IA » pourrait bien être le prochain phénomène de ce type.

***

Considérant que l'IA peut être définie comme une série de théories et de techniques utilisées pour créer des machines capables de simuler l'intelligence humaine[1]. Il s'agit d'une discipline scientifique mêlant les mathématiques statistiques et algorithmiques, les sciences informatiques et les sciences cognitives. L'IA symbolique est fondée sur la logique tandis que l'IA connexionniste s'appuie sur des réseaux de neurones artificiels.

Considérant que l'apprentissage automatique est un exemple d'IA connexionniste, tout comme l'apprentissage profond, lequel est un sous-ensemble de l'apprentissage automatique utilisant des neurones artificiels interconnectés et organisés en une multitude de couches. Plus le nombre de couches de neurones permettant l'apprentissage autonome est élevé, plus la complexité technologique du système augmente, ce qui rend celui-ci plus efficace mais aussi ses calculs moins explicables et moins traçables (apprentissage profond).

Considérant que l'apprentissage automatique peut rendre la technologie extrêmement puissante mais que le processus décisionnel dont celle-ci est dotée peut être si complexe qu'il fait penser à une « boîte noire ».

Notant que de nombreux systèmes d'IA utilisés pour prévenir la commission d'infractions, enquêter sur celles-ci, les détecter et les punir sont des systèmes d'apprentissage automatique. Ils procèdent en quelques nanosecondes à des calculs complexes de probabilités en s'appuyant sur des algorithmes d'auto-apprentissage. Pour atteindre les objectifs qui leur sont confiés, ils traitent d'immenses quantités de données et consomment énormément d'énergie. Certains de ces systèmes ont recours à l'apprentissage profond, par exemple pour la reconnaissance faciale.

***

Constatant que pour l'administration de la justice pénale, des systèmes d'IA sont utilisés pour prévenir ou détecter les infractions pénales et s'appuient à cet effet sur l'évaluation des risques (« police prédictive »).

Constatant que les systèmes d'IA sont aussi utilisés pour aider les procureurs et les juges à prendre des décisions, et plus précisément que l'expression « justice prédictive » désigne le fait : i) de prédire le comportement d'une personne, par exemple d'évaluer le risque que celle-ci s'enfuie pendant l'enquête préliminaire ou qu'elle récidive, afin qu'une décision puisse être prise à son égard, par exemple un placement en détention provisoire, une condamnation, une libération conditionnelle ou une condamnation avec sursis (la justice actuarielle, qui est de nos jours appuyée par l'IA) ; ii) d'utiliser l'IA pour procéder à une analyse statistique ultrarapide des décisions antérieures dans des affaires

[1] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

similaires et des dispositions législatives et réglementaires applicables (analyse quantitative du droit, ou « legaltech »).

Considérant que l'analyse quantitative du droit est révolutionnaire en ce sens qu'un calcul mathématique est censé venir étayer, voire supplanter, le raisonnement juridique.

Observant que le mot « prédictive », tel qu'il est employé dans les expressions « police prédictive » et « justice prédictive », est source de confusion car les systèmes d'IA calculent des probabilités mais ne prédisent pas l'avenir : ces probabilités sont fondées sur des corrélations et non sur la causalité. Ces calculs ont néanmoins un effet performatif sur les personnes, c'est-à-dire qu'ils peuvent les amener à prendre des décisions compatibles avec les résultats obtenus. Généralement parlant, l'ancrage scientifique de l'IA encourage ses utilisateurs à se fier aux probabilités que ses systèmes calculent et à s'en inspirer, car le « biais d'automatisation » est d'autant plus élevé que le système est paré d'une certaine aura scientifique.

Constatant que les systèmes d'IA contribuent à l'innovation en matière de recherche de preuves. Ils sont capables d'analyser rapidement des mégadonnées et d'en extraire les informations qui pourraient être utiles aux enquêteurs. Les systèmes d'IA peuvent déceler des corrélations invisibles à l'œil humain au sein des informations qu'ils traitent. Les diagrammes d'analyse du crime qu'ils dressent peuvent fournir aux enquêteurs des informations très précieuses et détaillées.

Constatant que les systèmes d'IA peuvent produire des informations qui serviront ensuite de preuves dans un procès pénal. Ils peuvent en particulier fournir des informations médico-légales en comparant les caractéristiques biométriques (par exemple, les images de visages pour la reconnaissance faciale), les fréquences de différentes voix (pour la reconnaissance vocale) et des échantillons d'ADN (pour le génotypage probabiliste).

Constatant, en dernier lieu, que les robots assistés par l'IA et les objets « intelligents » qui sont utilisés au quotidien dans divers domaines d'activité peuvent incidemment produire des indices ou des preuves susceptibles de s'avérer utiles pour établir des faits dans une affaire pénale.

***

Constatant que, malgré d'importants progrès ces quelques dernières années, les systèmes d'IA ne sont pas entièrement fiables. Des erreurs peuvent être commises en raison de la mauvaise qualité des données utilisées, de la manière dont les algorithmes sont programmés ou de l'existence de faux positifs/négatifs dans les corrélations. Il se peut donc que les probabilités calculées par un système d'IA soient inexactes.

Constatant que les résultats obtenus par les systèmes d'IA ne sont pas toujours totalement neutres. Pour pouvoir calculer des probabilités fiables, les systèmes d'IA sont non seulement tributaires de la qualité des données collectées et traitées, qui peuvent induire des biais, mais aussi de la manière dont les systèmes ont été entraînés (apprentissage non

supervisé ou supervisé). Étant donné qu'ils reproduisent des décisions humaines, les algorithmes d'auto-apprentissage sont influencés par les travers humains. Partant, il existe des algorithmes xénophobes[2], racistes, misogynes, etc.

Constatant que les systèmes d'IA posent des problèmes de transparence. Ce que l'on nomme « l'IA à boîte noire » est si opaque que même les spécialistes ne parviennent pas à déterminer comment les résultats sont obtenus. Même les experts scientifiques ne sont pas capables d'expliquer aux tribunaux, dans son intégralité, le raisonnement qu'un système a suivi.

Considérant que les systèmes d'IA qui sont employés dans le cadre de la justice pénale peuvent être mis au point par le secteur privé, et sont donc des produits à vendre qui doivent être rentables. Les entreprises qui les mettent au point se prévalent en règle générale de la protection du secret industriel pour refuser de révéler le code source de leur algorithme. Or, sans ce code source, le fonctionnement du système ne peut pas être dûment analysé.

***

Observant que tout ce qui est technologiquement possible n'est pas pour autant socialement souhaitable. Dans une démocratie, les choix politiques relatifs à la prévention des infractions pénales, à leur détection, aux enquêtes et aux poursuites les concernant doivent être traduits dans une loi ou une norme ayant la même force contraignante.

Réaffirmant que les droits humains doivent être pleinement protégés dans le cadre de la prévention des infractions pénales, de leur détection, des enquêtes et des poursuites les concernant, y compris lorsque des innovations technologiques sont utilisées à ces fins. Si le recours à l'IA soulève de fréquentes questions en matière de protection de la vie privée, de protection des données à caractère personnel et de non-discrimination, ce sont tous les textes protégeant les êtres humains, en particulier leur liberté et leur dignité, mais aussi les garanties d'un procès équitable et notamment la présomption d'innocence, qui sont potentiellement concernés par la menace liée à l'utilisation des technologies simulant l'intelligence humaine.

Considérant que les lois nationales et les normes juridiques internationales et/ou régionales sont en mesure de déterminer dans quelles conditions les innovations technologiques basées sur l'IA peuvent être autorisées à contribuer à l'administration de la justice pénale.

Réaffirmant que les normes éthiques sur lesquelles s'appuie le secteur privé n'ont pas la même force contraignante que la loi.

***

[2] https://www.amnesty.org/en/documents/eur35/4686/2021/en/

Tenant compte de :

- la Recommandation du Conseil sur l'intelligence artificielle, Organisation de coopération et de développement économiques, 22 mai 2019, C/MIN(2019)3/FINAL ;
- la Recommandation sur l'éthique de l'intelligence artificielle, de l'Organisation des Nations Unies pour l'éducation, la science et la culture, 23 novembre 2021, SHS/BIO/PI/2021/1 ;
- la Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement, Commission européenne pour l'efficacité de la justice (CEPEJ)), Strasbourg, 3-4 décembre 2018 ;
- la Résolution du Parlement européen du 6 octobre 2021 sur l'intelligence artificielle en droit pénal et son utilisation par les autorités policières et judiciaires dans les affaires pénales, document 2020/2016(INI) ;
- des résolutions du XIX[e] congrès international de droit pénal : Société de l'information et droit pénal, Rio de Janeiro, 2014.

***

Les participants au Colloque international de la Section III, qui s'est tenu à Buenos Aires, du 28 au 31 mars 2023, ont approuvé ce qui suit :

1.2 Résolutions

1. Le recours à des systèmes d'IA par les pouvoirs publics pour faciliter la prévention des infractions pénales, leur détection ou les enquêtes les concernant doit être préalablement autorisé par une loi ou une norme ayant la même force contraignante[3].

2. Les États doivent s'assurer que les décisions des autorités en matière de prévention et de détection d'un certain type d'infractions pénales et d'enquêtes sur celles-ci se fondent sur des critères définis en application d'un processus politique et démocratique et non sur l'hypothèse que le recours à la technologie de l'IA leur facilitera ces tâches.

3. Pour protéger la légitimité des activités des pouvoirs publics en matière de prévention des infractions pénales, de détection de celles-ci et d'enquêtes les concernant, les États qui souhaitent employer des systèmes d'IA doivent opter pour ceux dont le fonctionnement est totalement transparent, explicable et traçable (les IA à boîtes blanches). Ils doivent s'assurer à cet effet qu'aucune objection fondée sur la propriété intellectuelle ne puisse être formulée en réponse à des demandes liées à la transparence et ils devraient privilégier des systèmes à code ouvert mis à la disposition du public.

4. Les lois ou normes équivalentes relatives à l'utilisation de systèmes d'IA pour la prévention des infractions pénales, leur détection et les enquêtes les concernant doivent exiger que lesdits systèmes soient hautement fiables du point de vue technologique. Une

[3] Ci-après, l'expression « norme ayant la même force contraignante » sera abrégée en « normes équivalentes ».

réglementation suffisamment précise, exigeant que des entités externes aux développeurs et fournisseurs de systèmes d'IA et indépendantes d'eux procèdent aux vérifications et évaluations voulues, doit limiter dans toute la mesure du possible le risque de biais ou de toute autre forme de discrimination dans le cadre de l'apprentissage automatique ainsi que le risque d'erreurs d'encodage et d'autres dysfonctionnements technologiques.

5. Les lois ou normes équivalentes doivent exiger que les systèmes d'IA utilisés pour faciliter la prévention des infractions pénales, leur détection et les enquêtes les concernant soient pleinement accessibles, vérifiables et contrôlables par les autorités qui s'en servent et par celles qui sont chargées de leurs vérifications et évaluations.

6. Les lois ou normes équivalentes autorisant l'utilisation de systèmes d'IA pour faciliter la prévention des infractions pénales, leur détection et les enquêtes les concernant doivent exiger que les données d'entraînement soient de grande qualité et hautement représentatives.

S'agissant des données contenues dans les fichiers de police ou de la justice, les lois ou normes équivalentes doivent instaurer un système permettant de s'assurer que ces données sont exactes et à jour et que leur utilisation ne porte pas atteinte à la présomption d'innocence. La présomption d'innocence interdit formellement la rétention et l'utilisation de données recueillies en conséquence d'une évaluation prédictive lorsque la culpabilité n'a pas été établie par la suite, à la seule exception du cas dans lequel les données présentent un intérêt à l'égard d'un autre suspect.

S'agissant des autres données, en particulier celles qui sont accessibles sur les médias sociaux, les lois ou normes équivalentes doivent exiger que lesdites données soient utilisées dans le respect du droit à la vie privée et des règles de protection des données à caractère personnel. Des vérifications appropriées et indépendantes de la police et des institutions judiciaires doivent être effectuées.

De manière générale, des lois ou normes équivalentes doivent poserdes exigences strictes quant à la vérification de la fiabilité de toutes les données utilisées par les systèmes d'IA aux fins de la détection des infractions pénales, de leur prévention et des enquêtes les concernant.

7. Les lois ou normes équivalentes doivent exiger, préalablement à l'utilisation de systèmes d'IA reposant sur des algorithmes d'auto-apprentissage pour la prévention des infractions pénales, leur détection ou pour les enquêtes les concernant, que lesdits algorithmes soient mis au point, entraînés, testés et déployés sous supervision humaine (apprentissage automatique avec un « humain dans la boucle »).

Ces lois et normes équivalentes doivent exiger qu'une évaluation humaine soit réalisée avant l'adoption de toute mesure visant à prévenir et détecter des infractions pénales ou à enquêter à leur sujet sur le fondement de probabilités calculées par un système d'IA.

8. Les États et les autorités répressives doivent non seulement s'assurer que leur personnel utilisant l'IA pour prévenir des infractions pénales, les détecter ou enquêter à leur sujet reçoive une formation pratique sur la bonne utilisation des systèmes d'IA concernés ainsi qu'une formation sur le risque d'erreur et les biais, mais également s'assurer que ledit personnel soit suffisamment informé des dangers que l'IA représente pour les droits humains.

9. Les autorités internationales, régionales, nationales ou locales doivent mettre en place des organes indépendants pour homologuer la qualité des systèmes d'IA censés être utilisés aux fins de la prévention des infractions pénales, de leur détection ou des enquêtes les concernant. Une technologie d'IA qu'il n'est pas possible de faire fonctionner ou de superviser de manière transparente pour des raisons entre autres liées aux droits de propriété intellectuelle ne doit pas être homologuée.

Le secteur privé devrait s'organiser ou s'unir pour élaborer des labels de qualité des systèmes d'IA qui permettraient de créer un cercle vertueux pour ces produits afin que les autorités soient mieux à même de déterminer quels systèmes d'IA correspondent le mieux à leurs besoins lorsqu'elles cherchent à prévenir des infractions pénales, à les détecter ou à enquêter à leur sujet.

10. Tous les droits humains doivent être protégés lorsque des systèmes d'IA sont employés aux fins de la prévention des infractions pénales, de leur détection ou des enquêtes les concernant. Les États ainsi que les organismes régionaux et internationaux doivent veiller à ce que des sanctions effectives, proportionnées et dissuasives soient imposées en cas de violation de ces droits.

Les lois ou normes équivalentes doivent expressément prévoir que lorsque la violation de droits humains est causée par le dysfonctionnement technologique d'un système d'IA, l'entreprise qui a créé ce dernier soit tenue responsable pour faute ou pour négligence ou soumise à une responsabilité de plein droit du fait de la défectuosité dudit produit. Elles doivent également prévoir l'obligation de mener des enquêtes pour déterminer les causes de la violation.

11. Toutes les résolutions énoncées ici présent sont également applicables à la prévention des infractions administratives, à leur détection, aux enquêtes les concernant et à la sanction desdites infractions par les autorités compétentes.

1.3 Résolutions portant expressément sur la « police prédictive »

12. Les États et les organismes régionaux et internationaux compétents en matière de droits humains doivent s'assurer que l'utilisation des systèmes d'IA dans la prévention et la détection des infractions pénales ne se solde pas par une surveillance de masse, laquelle entraînerait une réduction disproportionnée des libertés individuelles (liberté de mouvement, liberté d'expression, liberté de réunion, liberté d'association et liberté de religion).

Les États et les autorités locales doivent en particulier interdire l'utilisation de systèmes d'IA pour l'identification de personnes physiques, à distance et à l'aide de leurs données biométriques, dans des espaces accessibles au public, ainsi que toute autre utilisation des systèmes d'IA à des fins de surveillance de masse.

Les États sont exhortés de faire preuve de davantage de transparence quant à leur utilisation de systèmes de reconnaissance automatique des plaques d'immatriculation dans les espaces accessibles au public. Lorsque ces systèmes permettent de photographier non seulement la plaque minéralogique mais également toute personne présente dans le véhicule, cette dernière possibilité doit être expressément autorisée par la loi. Il doit être interdit que les données collectées à l'aide de ces photographies soient analysées par un dispositif de reconnaissance faciale à des fins de « police prédictive », sauf dans le cadre d'une enquête précise et si une disposition légale l'autorise.

13. Les États doivent déterminer – ou confier cette tâche à des organismes de recherche indépendant – si l'utilisation de systèmes d'IA pour la prévention des infractions pénales contribue à la réduction du nombre d'infractions commises et, si tel est le cas, dans quelle proportion.

14. Les États doivent s'assurer que le coût financier des systèmes d'IA et de leur maintenance n'entraîne pas une réduction de l'enveloppe allouée aux services publics de prévention des infractions qui travaillent sur les *causes* de celles-ci (à des fins de soutien psychologique, de soutien social, de formation et de soutien à l'emploi).

15. Les lois et normes équivalentes doivent interdire purement et simplement l'utilisation en tant que preuves à charge de données recueillies par un système d'IA aux fins de prévention des infractions, ce lorsqu'aucun élément concret ne permet de soupçonner qu'une infraction a été commise et que, par conséquent, les données ont été recueillies en dehors du champ d'application de la règlementation régissant les enquêtes pénales. Si des données recueillies par un système d'IA à des fins de prévention des infractions sont utilisées comme point de départ d'une enquête pénale (soupçon), l'autorité judiciaire compétente doit en être informée. Les données doivent être désignées en tant que telles et le recours à des systèmes d'IA doit être mentionné dans le dossier de la procédure.

1.4 Résolutions portant expressément sur la « justice prédictive »

16. Les lois et normes équivalentes doivent interdire formellement que des systèmes d'IA soient utilisés pour déterminer des peines dans le cadre d'une démarche de justice actuarielle.

Condamner une personne ou aggraver sa peine en fonction de la probabilité que celle-ci commette une infraction pénale à l'avenir revient à condamner cette personne en partie pour une infraction pénale qui n'a pas encore été commise, ce qui porte atteinte à la dignité humaine, à la liberté individuelle et aux principes fondamentaux de la justice pénale.

L'utilisation d'outils d'évaluation des risques par l'IA doit être interdite lorsque des mesures de sûreté graves, par exemple la détention, sont en jeu. Lorsque les États autorisent l'usage de tels outils pour des mesures moins graves, il faut alors que la loi l'autorise expressément et prévoie des garanties procédurales suffisantes. Les probabilités calculées par l'IA ne peuvent toutefois pas constituer le seul fondement sur lequel repose une décision.

17. Les États qui souhaitent que l'IA soit utilisée pour assister les procureurs et/ou les juges en procédant à une analyse quantitative du droit en amont de la prise de décisions dans une affaire pénale doivent limiter le recours à cette technologie à des infractions mineures représentant un volume élevé de dossiers.

18. Avant de décider d'utiliser l'IA pour faciliter la gestion d'un volume élevé de dossiers concernant des infractions mineures, les États doivent vérifier s'il ne serait pas judicieux, en vertu du principe de l'*ultima ratio*, de dépénaliser les comportements qui ont généré ces dossiers.

19. Les lois et normes équivalentes doivent interdire le recours à l'analyse quantitative du droit pour aider les juges à se prononcer sur la culpabilité.

20. Les lois et normes équivalentes doivent interdire le recours à l'analyse quantitative du droit pour aider les juges à prononcer une condamnation. La décision de punir une personne et le choix de la peine relèvent des êtres humains, sans quoi la justice pourrait être déshumanisée et la dignité humaine mise en péril.

21. Les lois et normes équivalentes doivent interdire que l'analyse quantitative du droit soit employée pour aider les juges à prendre des décisions en matière pénale avant jugement qui impliquent des mesures coercitives.

22. Les États doivent s'assurer que les décisions prises à l'aide de l'analyse quantitative du droit ne portent pas atteinte au droit d'accès à un juge humain.

23. Les lois et normes équivalentes doivent interdire qu'une décision soit prise à l'aide de l'analyse quantitative du droit sauf si la personne visée par ladite décision peut introduire un recours contre celle-ci. La décision rendue sur recours ne pourra pas être fondée uniquement sur une l'analyse quantitative du droit.

1.5 Résolutions portant expressément sur les preuves recueillies et/ou produites par l'IA

24. Les lois et normes équivalentes concernant l'extraction de données aux fins d'une analyse par un système d'IA doivent exiger qu'avant de demander à une personne le code d'accès à ses logiciels ou matériels dont des données pourraient être extraites, l'autorité saisissant les données informe la personne concernée de son droit de ne pas témoigner contre elle-même.

25. Les lois et normes équivalentes sur l'analyse du crime doivent préciser que les diagrammes d'analyse du crime qui sont produits par les systèmes d'IA n'ont pas de valeur probante mais peuvent venir éclairer le déroulement d'une enquête.

26. Les lois et normes équivalentes relatives à l'usage de systèmes d'IA pour la collecte de preuves ou la production d'informations aux fins de justice pénale doivent indiquer clairement que les résultats obtenus à l'aide desdits systèmes ne sont que des probabilités. Elles doivent exiger que les jugements rendus à l'aide de probabilités précisent non seulement quelles probabilités calculées par un système d'IA ont été utilisées mais également le taux d'erreur du système tel qu'il a été calculé par l'organisme d'homologation qui l'a évalué.

27. Les États et les autorités judiciaires doivent s'assurer que l'usage de probabilités calculées par l'IA n'entraîne pas une diminution du niveau de preuve exigé dans les procédures pénales.

28. Les lois et normes équivalentes relatives à l'usage de systèmes d'IA pour la collecte de preuves ou la production d'informations aux fins de justice pénale doivent interdire que les probabilités calculées par des systèmes d'IA dont il n'est pas possible d'expliquer pleinement le fonctionnement (IA à boîtes noires) soient considérées comme des preuves.

29. Les lois et normes équivalentes relatives à l'usage de systèmes d'IA pour la collecte de preuves ou la production d'informations aux fins de justice pénale doivent exiger, conformément au droit à un procès contradictoire, que s'il est fait usage de données recueillies ou produites par un système d'IA, toutes les parties en soient informées. Les données doivent être désignées en tant que telles et le recours à des systèmes d'IA doit être mentionné dans le dossier du procès.

Les lois et normes équivalentes doivent exiger que lorsqu'une partie produit des probabilités calculées par IA, l'autre partie puisse les récuser.

30. Les lois et normes équivalentes doivent énoncer le principe selon lequel la partie produisant en justice des probabilités doit en toutes circonstances préciser la manière dont le système d'IA concerné fonctionne et sur quelles données il fonde son calcul de probabilités.

31. Les lois et normes équivalentes relatives à l'usage de systèmes d'IA pour la collecte de preuves ou la production d'informations aux fins de justice pénale doivent prévoir, conformément aux droits de la défense, que toute personne accusée d'avoir commis une infraction sur la base de probabilités produites en tant que preuves puisse obtenir, afin qu'un expert les analyse, le code source et les données d'entraînement du système d'IA qui a calculé lesdites probabilités. Le secret industriel ne doit pas pouvoir être invoqué pour empiéter sur les droits de la défense.

32. Compte tenu du coût élevé de l'expertise d'un système d'IA, les États doivent s'assurer que toute personne accusée d'une infraction sur la base d'une probabilité calculée par

un système d'IA ait accès non seulement à une aide juridique effective mais également à une aide financière pour ladite expertise.

Section IV: Perspectives internationales sur l'IA : les enjeux pour la coopération judiciaire, le droit international humanitaire et le droit pénal

1.6 Préambule

Considérant que l'intelligence artificielle (IA) fait partie intégrante du processus de prise de décisions lié aux systèmes d'armes autonomes (SAA), que ces armes ont déjà atteint un niveau de développement avancé en cette troisième décennie du 21e siècle, et qu'elles sont déjà largement employées par de nombreux États et par des acteurs non étatiques.

Considérant qu'une multitude de décisions en matière de surveillance et de ciblage qui étaient habituellement prises ou étayées par des humains commencent à être automatisées en raison de l'usage des SAA, ce qui pose d'importantes questions d'ordre juridique, moral et éthique.

Soulignant les préoccupations grandissantes que suscite la nuisibilité des usages malveillants ou imprudents des SAA ;

Reconnaissant qu'il faut étudier si les États et la communauté internationale apportent une réponse juridique adéquate aux défis que pose l'IA ou s'il faut procéder à des réformes et adaptations, soit en modifiant expressément certaines dispositions soit en créant de nouvelles formes d'incrimination ;

Reconnaissant qu'il faut que les SAA soient conçus et mis au point dans le respect du droit international.

***

Soulignant que les SAA peuvent être définis comme tout système d'armes ayant pour caractéristique de posséder une autonomie dans ses fonctions critiques : c'est-à-dire de pouvoir sélectionner (rechercher, détecter, identifier, localiser ou choisir) et attaquer des cibles (utiliser la force contre elles, les neutraliser, les endommager ou les détruire) sans intervention humaine[4] ;

Notant toutefois que cette définition des SAA ne fait pas l'unanimité et que dans d'autres structures internationales, régionales et nationales les SAA et le contrôle humain significatif de leur utilisation (« l'humain dans la boucle ») peuvent être interprétés différemment ;

[4] CICR, *Position du CICR sur les systèmes d'armes autonomes*, document communiqué lors de la réunion du Groupe d'experts gouvernementaux (GGE) de la Convention sur certaines armes classiques (CCAC) au sujet des systèmes d'armes létaux autonomes (SALA), 11 avril 2016, https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system (en anglais uniquement)

***

Notant que d'autres structures internationales, régionales et nationales peuvent en outre choisir de différencier les SAA et les SALA ainsi que les SAA et les systèmes d'armes automatisés ;

Considérant que le recours aux SAA peut causer des dommages considérables aux humains et aux sociétés et qu'il a des répercussions d'ordre juridique et éthique aussi bien en ce qui concerne le *jus ad bellum* et le *jus in bello*, qu'en termes de violations graves des droits humains, de violations du *jus cogens* et que par rapport au noyau dur des crimes internationaux ;

Considérant que le recours aux SAA peut faire pencher l'opinion publique et les politiques publiques en faveur de l'usage de la force à cause de l'impression que les SAA atténuent les risques de destruction et les risques de tuer ou blesser des soldats et autres personnes impliquées ;

Considérant que le recours aux SAA peut avoir des incidences négatives sur le respect des principes fondamentaux du *jus in bello*, à savoir les principes de distinction, de proportionnalité et de précaution ;

Considérant qu'en retirant à l'élément humain son rôle dans la prise de décisions ou en réduisant celui-ci, l'utilisation des SAA peut contribuer à la hausse du nombre de décès à cause de l'absence de sentiments humains comme la peur et la compassion, lesquels peuvent concourir à la réduction du nombre de décès ;

Considérant que l'utilisation des SAA peut causer des destructions et des dommages collatéraux considérables ;

Considérant que les quatre grands crimes internationaux peuvent être commis moyennant l'utilisation de SAA et qu'une telle utilisation peut soulever de nouvelles questions, graves, en ce qui concerne l'attribution de la responsabilité pénale mais aussi, notamment, la responsabilité du supérieur hiérarchique.

***

Considérant que l'utilisation des SAA peut soulever des questions de compétence car elle peut être transterritoriale ;

Constatant qu'il faut adopter des règles internationales relatives aux SAA, en ce compris une norme juridique contraignante interdisant la conception, la mise au point, la production et l'utilisation de systèmes d'armes entièrement autonomes (c'est-à-dire sans contrôle humain significatif) ;

Encourageant tous les États à entamer officiellement des négociations dans le cadre des Nations Unies (sous l'égide de la Convention sur l'interdiction ou la limitation de l'em-

ploi de certaines armes classiques et/ou d'autres conventions, entités, agences et institutions concernées des Nations Unies) pour élaborer des règles ou des lignes directrices relatives à la conception, à la mise au point, à la production et à l'utilisation des SAA ;

Constatant que cela renforce également la nécessité d'une approche mondiale des SAA.

***

1.7 Tenant compte :

- du rapport de 2016 des Nations Unies : *Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*[5] ;

- de la Convention des Nations Unies sur l'interdiction ou la limitation de l'emploi de certaines armes classiques qui peuvent être considérées comme produisant des effets traumatiques excessifs ou comme frappant sans discrimination, telle qu'elle a été modifiée le 21 décembre 2001[6] ;

- de la Résolution du Parlement européen du 12 septembre 2018 sur les systèmes d'armes autonomes[7] ;

- de la Conférence régionale Amérique latine-Caraïbes sur l'impact social et humanitaire des armes autonomes et du Communiqué de ladite conférence, des 23 et 24 février 2023 ;

- du rapport du Comité international de la Croix-Rouge (CICR) intitulé : « *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* », Genève, septembre 2016[8] ;

- de la déclaration du CICR intitulée : « *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)* », 13-17 avril 2015, Genève[9] ;

- du XX[e] Congrès de l'AIDP, tenu à Rome du 13 au 16 novembre 2019 sur le thème « justice pénale et droit des sociétés »[10].

---

[5] https://documents.un.org/doc/undoc/gen/g16/117/16/pdf/g1611716.pdf (en anglais uniquement).

[6] https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/#:~:text=It%20was%20adopted%20on%2010,or%20to%20affect%20civilians%20indiscriminately

[7] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_FR.html

[8] CICR (en anglais uniquement), *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, CICR, Genève, septembre 2016, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems

[9] CICR (2015) (en anglais uniquement) *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 13-17 avril 2015, Genève*, https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS

[10] https://www.penal.org/fr/aidp-xx-international-congress-penal-law-%E2%80%9Ccriminal-justice-and-corporate- business%E2%80%9D-rome-13-16-0

Les participants au Colloque international de la Section IV, qui s'est tenu à Opatija, Croatie, du 7 au 8 décembre 2023, ont approuvé ce qui suit :

Résolutions

1) Afin d'empêcher et d'atténuer les dommages liés aux SAA, il est nécessaire que les législateurs, aux échelons international, régional et national, et les autres autorités compétentes, élaborent une définition approfondie de ce que sont les SAA et qu'ils établissent des réglementations régissant la conception, la mise au point, la production et l'utilisation des SAA.

2) L'emploi des SAA doit être préalablement réglementé par une loi ou une norme de force contraignante similaire[11].

3) Les États sont exhortés à faire preuve de davantage de transparence quant à l'utilisation qu'ils font des SAA.

4) Les États doivent s'assurer que la décision par les autorités d'utiliser des SAA est fondée sur des critères transparents faisant l'objet d'un contrôle public. Les États qui utilisent des SAA doivent opter pour les systèmes dont le fonctionnement est pleinement transparent, explicable et traçable (IA boîtes blanches).

5) Les lois ou les normes équivalentes relatives à l'utilisation des SAA doivent exiger que ces systèmes soient hautement fiables sur le plan technologique. Il faut qu'une réglementation suffisamment précise, exigeant que des entités externes aux développeurs de SAA et indépendantes d'eux procèdent aux vérifications et évaluations voulues, limite dans toute la mesure du possible le risque de biais ou de toute forme de discrimination dans le cadre de l'apprentissage automatique, d'erreurs d'encodage et d'autres dysfonctionnements technologiques.

6) Les lois ou normes équivalentes doivent exiger qu'avant qu'un SAA basé sur des algorithmes d'auto-apprentissage soit utilisé à des fins de surveillance, choix de cibles ou à toute autre fin, les algorithmes soient mis au point, entraînés, testés et utilisés sous supervision humaine (principe des « humains dans la boucle »). Ces lois et normes équivalentes doivent exiger une évaluation humaine avant que la moindre mesure soit prise en vue de l'utilisation d'un SAA.

7) Les États et les autorités répressives doivent non seulement s'assurer que leur personnel chargé de faire fonctionner des SAA reçoit une formation sur la bonne utilisation de

[11] L'article 36 du Protocole additionnel 1 aux Conventions de Genève exige déjà des Hautes Parties contractantes qu'elles déterminent si la mise au point ou l'emploi d'une nouvelle arme seraient interdits par les dispositions dudit Protocole ou par toute autre règle du droit international applicable. Protocole additionnel aux Conventions de Genève du 12 août 1949 relatif à la protection des victimes des conflits armés internationaux (Protocole I), article 36. https://www.ohchr.org/fr/instruments-mechanisms/ instruments/ protocol-additional-geneva-conventions-12-august-1949-and.

ces systèmes ainsi qu'une formation sur le risque d'erreur et de biais, mais également s'assurer que ledit personnel est dûment informé des dangers que les SAA représentent pour les droits humains.

8) Les lois ou normes équivalentes doivent expressément prévoir qu'en cas de violation de droits humains causée par tout dysfonctionnement d'un SAA, l'entreprise qui a conçu ou fabriqué le système soit tenue pénalement responsable pour faute ou pour négligence[12] et civilement responsable du fait de la défectuosité du produit. Ces lois ou normes équivalentes doivent aussi prévoir l'obligation de mener des enquêtes pour déterminer les causes de la violation.

9) Doivent être en vigueur dans chaque État qui choisit d'utiliser des SAA des lois pénales prévoyant que les personnes physiques utilisant des SAA de façon non conforme aux normes juridiques internationales et nationales applicables verront leur responsabilité pénale individuelle engagée devant les juridictions nationales compétentes[13].

10) Les États et les organismes régionaux et internationaux compétents en matière de droits humains doivent s'assurer que l'utilisation des SAA ne donne pas lieu à des violations graves des droits humains, à des violations du *jus cogens* ou autres violations du droit international applicable et des normes du droit international humanitaire.

11) Les États doivent s'assurer qu'ils ont pleinement transposé leurs obligations juridiques de droit international dans leurs droits nationaux respectifs, que ce soit en adoptant des textes de loi, des règlements, des décrets, un code militaire ou autre type de normes juridiques nationales ayant force contraignante. Le droit national doit prévoir différents modes de responsabilité, en ce compris la responsabilité du supérieur hiérarchique.

12) Les États devraient créer des organes de recherche indépendants chargés d'examiner en permanence si l'utilisation des SAA est conforme à l'ensemble des réglementations et normes juridiques internationales et nationales applicables.

13) Il est nécessaire de déterminer et de définir des modes précis d'attribution de la responsabilité pénale aux personnes (aussi bien physiques que morales) qui sont responsables de la mise au point des SAA[14]. Il conviendrait de ne pas exclure la responsabilité des personnes, physiques ou morales, ayant apporté une contribution dans la chaîne de

[12] XXIe Congrès de l'AIDP : intelligence artificielle et justice pénale, Section 1 : les catégories traditionnelles de droit pénal et l'IA, résolutions approuvées au colloque de Syracuse, 15-17 septembre 2022 ; XXIe Congrès de l'AIDP : intelligence artificielle et justice pénale, Section 2 : le droit pénal et l'incrimination face aux défis qu'engendre l'IA, résolutions approuvées au colloque de Bucarest, 13-17 juin 2023.
[13] XXIe Congrès de l'AIDP : intelligence artificielle et justice pénale, Section 1 : les catégories traditionnelles de droit pénal et l'IA, résolutions approuvées au colloque de Syracuse, 15-17 septembre 2022.
[14] XXIe Congrès de l'AIDP : intelligence artificielle et justice pénale, Section 1 : les catégories traditionnelles de droit pénal et l'IA, résolutions approuvées au colloque de Syracuse, 15-17 septembre 2022.

causalité du préjudice, qu'il s'agisse des concepteurs, des programmeurs, des producteurs, des vendeurs, des distributeurs ou des utilisateurs finaux eux-mêmes des SAA.

14) Les États devraient s'efforcer de favoriser la coopération entre les experts politiques et universitaires du droit international et les personnes appartenant à l'armée, aux services répressifs ou aux services de renseignement qui sont chargés de mettre au point et/ou de déployer les SAA. Il faudrait mener cette coopération dans l'objectif de s'assurer que l'utilisation des SAA est conforme aux normes de droit international applicables ainsi qu'aux lois et réglementations nationales applicables.

15) Les États doivent prévoir l'application extraterritoriale de leur droit interne dans les cas où des SAA sont déployés ou autrement utilisés à l'étranger par eux ou par des entités placées sous leur contrôle.

16) Les États doivent adopter des lois relatives au conflit de lois ou au conflit de compétence afin que le droit interne puisse s'appliquer dans les cas où des SAA sont déployés ou utilisés à l'étranger. Ces lois pourraient être élaborées aux échelons national, régional et/ou international. Elles doivent permettre d'engager la responsabilité pénale (ainsi que la responsabilité civile) des personnes physiques ou morales qui sont responsables de l'utilisation abusive ou du dysfonctionnement des SAA hors du territoire de l'État concerné.

17) Les États doivent établir des modes de coopération judiciaire en matière pénale, en particulier afin de collecter des preuves relatives aux infractions causées par des SAA. Ces modes de coopération judiciaire pourraient s'avérer particulièrement nécessaires pour les États qui n'ont pas encore conclu d'accords bilatéraux, régionaux ou internationaux à cet égard.

18) Les États doivent s'assurer qu'ils disposent de mécanismes d'extradition afin que les personnes responsables des infractions causées par les SAA puissent être dûment poursuivies et punies.

XXI CONGRESO INTERNACIONAL DE DERECHO PENAL
(París, 25-28 junio 2024)

Tema: Inteligencia Artificial y Justicia Penal

1. Las categorías tradicionales del Derecho Penal y la inteligencia artificial: ¿Crisis o Palingenesia?

2. La criminalización de los delitos relacionados con la IA

3. La IA y la administración de justicia penal: 'policía predictiva', 'justicia predictiva' y derecho probatorio

4. Perspectivas internacionales sobre la IA: Desafíos para la Cooperación Judicial y el Derecho Penal/ Humanitario Internacional

Sección I: Las categorías tradicionales del Derecho Penal y la inteligencia artificial: ¿Crisis o Palingenesia?

Preámbulo

Considerando que

- la llegada de sistemas de Inteligencia Artificial (IA) con distinto grado de autonomía apoya y sustituye muchas actividades humanas;

- los sistemas de IA podrían representar un beneficio real para la sociedad en su conjunto y para los cuerpos y fuerzas de seguridad, concretamente en lo que se refiere a las investigaciones de infracciones penales;

- los sistemas de IA son cada vez más autónomos y su funcionamiento puede resultar imprevisible para quienes los diseñan, programan, producen, distribuyen y utilizan;

Observando que

- los ámbitos de aplicación de los sistemas de IA son considerablemente amplios, y los hechos ilícitos relacionados con su implementación podrían lesionar diferentes intereses, bienes jurídicos y derechos fundamentales;

- los sistemas de IA también pueden desempeñar un papel cada vez más importante como 'instrumento' para la comisión de actos delictivos, y podrían convertirse en el factor facilitador de la aparición de nuevos delitos;

Prestando especial atención a

- la creciente delegación total o parcial de decisiones en los sistemas de IA en diferentes ámbitos de actividad, que plantea la cuestión de la responsabilidad de las personas físicas o jurídicas por los daños causados por el funcionamiento autónomo de dichos sistemas;

- la autonomía de los sistemas de IA, que también ha generado un debate en el mundo académico relacionado con la posibilidad de considerarlos autores de un delito;

Teniendo en cuenta

- la importancia de las reacciones apropiadas que el Derecho Penal debe asegurar para prevenir y castigar los delitos cometidos por, a través o contra los sistemas de IA;

- la gravedad de los daños y de los riesgos relacionados con las aplicaciones de la IA;

- los principios fundamentales que deben garantizarse al establecer y aplicar sanciones penales (incluidas también las sanciones punitivas en un sentido más amplio, aplicables a las personas jurídicas), como el principio de legalidad y el principio de culpabilidad, que es una expresión necesaria de la personalidad de la responsabilidad penal;

- conscientes de que las categorías tradicionales del Derecho penal y los modelos de responsabilidad penal deben ser tomados en consideración y, en caso necesario, adaptados a los nuevos requisitos de protección;

Teniendo en cuenta

- las 'Directrices éticas para una IA confiable' presentadas a la Comisión Europea el 8 de abril de 2019 por el Grupo de Expertos de Alto Nivel, y otras recomendaciones significativas de otros organismos internacionales (por ejemplo, el Estudio de viabilidad sobre un futuro instrumento del Consejo de Europa sobre Inteligencia Artificial y Derecho Penal del Comité Europeo sobre Problemas de Delincuencia del Consejo de Europa, 4 de septiembre de 2020);

- la propuesta de Reglamento europeo sobre Inteligencia Artificial (la llamada Ley sobre IA); los trabajos del Comité sobre Inteligencia Artificial del Consejo de Europa; las actividades de las Naciones Unidas sobre Inteligencia Artificial;

- las recomendaciones del XIV Congreso Internacional (Viena, 1989) sobre los problemas jurídicos y prácticos que plantea la diferencia entre el Derecho penal y el Derecho penal administrativo, las del XVIII Congreso Internacional (Estambul, 2009), sobre la incriminación de la preparación y participación en un delito, y las del XIX Congreso Internacional (Río de Janeiro, 2014), sobre Sociedad de la Información y Derecho Penal,

Los participantes en el Coloquio Internacional de la Sección I, celebrado en Siracusa, del 14 al 17 de septiembre de 2022, aprueban lo siguiente:

Recomendaciones

I. Sobre el concepto de Inteligencia Artificial y la atribución de personalidad jurídica a sistemas de IA con distinto grado de autonomía

1. La creciente y compleja evolución de la IA pone de manifiesto que este concepto engloba múltiples sistemas algorítmicos y robóticos que interactúan con el entorno, desarrollados con diversas técnicas (como el *Machine Learning*), para la persecución de objetivos definidos por el ser humano. Por lo tanto, no es conveniente ofrecer una definición general de IA a efectos del Derecho penal.

2. Sin embargo, dado que los sistemas de IA pueden ser nocivos en diversos ámbitos (por ejemplo, vehículos autoconducidos, sistemas robóticos en medicina, sistema de negociación con IA o de gestión logística), la protección de bienes jurídicos y derechos fundamentales también en el ámbito del Derecho Penal no puede dejar de tener en cuenta las características específicas de los distintos sistemas de IA con diferente grado de autonomía, así como las definiciones jurídicas aportadas por fuentes extrapenales en cada uno de los sectores específicos.

3. En el estado actual de las cosas, no existe fundamento normativo ni coherencia en cuanto a las funciones del castigo penal en el reconocimiento de personalidad jurídica a los sistemas de Inteligencia Artificial con distinto grado de autonomía.

4. Por un lado, hay una diferencia ontológica respecto de los agentes humanos. Los sistemas de IA carecen de conciencia a la hora de elegir y evaluar las posibles soluciones a un problema o dilema, teniendo igualmente en cuenta el contexto de relaciones y oportunidades sociales y éticas, con la flexibilidad y capacidad necesarias para adaptarse a situaciones y condiciones incluso contingentes o sobrevenidas.

5. Por otro lado, las sanciones punitivas a tales sistemas y agentes tecnológicos no responderían a los fines y funciones de la sanción penal, pues el efecto de la amenaza de la pena y su aplicación se verían vaciados por la ausencia de autoconciencia de su propia existencia en el pasado, presente y futuro y, sobre todo, por la ausencia de autodeterminación voluntaria, de modo que, aun excluyendo la función retributiva, ni siquiera serían viables las de prevención especial y general.

II. Sobre la necesidad de regulación, normas y obligaciones extrapenales

6. Para prevenir y reducir los daños relacionados con la IA, antes o, al menos, paralelamente a las reformas del Derecho penal, sería necesario que los legisladores y las autoridades competentes internacionales, regionales y nacionales definieran plenamente la regulación de los diversos sectores en los que se implantan los sistemas de IA (como los mencionados anteriormente de vehículos autoconducidos, robots sanitarios y quirúrgicos, armas autónomas, etc.). Deben regularse las normas técnicas, las

características estructurales y las condiciones de funcionamiento de los sistemas de IA y sus componentes.

7. Dichas regulaciones, que han de contemplar desde las fases de diseño, producción, distribución y venta hasta el uso real de los sistemas de IA, también deben prever requisitos concretos de adaptación en caso de señales de alarma o alerta, como condición previa para abordar los daños relacionados con la IA a través del derecho punitivo.

8. Estas normativas podrían prever procedimientos interdictales, como los ya previstos en ámbitos de riesgos complejos (por ejemplo, la salud y la seguridad en el lugar de trabajo y la protección del medio ambiente), cuya violación o incumplimiento puedan sancionarse con arreglo al principio de *ultima rat*io.

III. Sobre la necesidad de la protección penal de los bienes jurídicos

9. Es necesario reconocer la importancia esencial de una intervención razonable y proporcionada del Derecho Penal en sentido amplio en la prevención y sanción de los daños y peligros que para los intereses, bienes jurídicos y derechos fundamentales pueden generar los sistemas de IA, dado que los mismos hechos, de ser realizados por personas físicas y jurídicas, según las categorías tradicionales del Derecho Penal, podrían ser constitutivos de infracción penal. Por tanto, no pueden quedar impunes simplemente por el hecho de su realización por, a través de o contra los citados sistemas.

10. Es necesario identificar y definir modelos específicos de atribución de responsabilidad a las personas (tanto físicas como jurídicas) que se encuentran 'detrás' de los sistemas de IA (es decir, a los actores de las distintas fases del ciclo de vida de los sistemas de IA, como diseñadores, proveedores, importadores, distribuidores, usuarios, etc.), empezando por los propietarios y quienes deciden su uso concreto, en función de su interés y su beneficio, y a los que, por tanto, debe exigirse responsabilidad jurídica, también desde una perspectiva 'punitiva' - no sólo desde el Derecho Penal-.

11. La responsabilidad de las personas descritas en el punto anterior no ha de excluir la de otras personas (físicas o jurídicas) que contribuyan a la cadena causal del daño: desde el diseñador, programador, productor, vendedor, distribuidor hasta los propios usuarios finales de los sistemas.

12. En particular, debe distinguirse entre:

a. Sistemas de IA utilizados en actividades ilícitas: en este ámbito se producirán principalmente conductas dolosas, que plantean menos problemas en cuanto a la atribución de responsabilidad penal, dado que los sistemas de IA no difieren conceptualmente de otros instrumentos y medios para cometer un delito.

No obstante, deben abordarse dos cuestiones

a.1. en caso de resultados derivados del funcionamiento del sistema desviados respecto de la actividad ilícita prevista, deben aplicarse los principios tradicionales de *aberratio*

*ictu*s y *aberratio delicti*. La mera diversidad material del objeto dañado no debe suponer una excusa si sus características no son relevantes para la configuración del ilícito penal (por ejemplo, matar a una persona en lugar de a otra no es relevante para la realización del delito de asesinato, cuando es intencionado por el agente). En su lugar, debería preferirse basar la responsabilidad penal (por un delito distinto del previsto) en la posibilidad de previsión concreta de ese desarrollo diferente de la acción puesta en marcha por el sistema de IA, aplicando los principios de la responsabilidad basada en la negligencia (como se expone *infr*a IV);

a.2. dado que los sistemas de IA pueden utilizarse para conductas especialmente dañinas o peligrosas, pueden amplificar y agravar el daño causado (como ocurre con las TIC). En efecto, las consecuencias pueden quedar muy alejadas de las acciones que les dieron origen, haciendo más difícil la intervención *post factum* para evitar o al menos detener o reducir los hechos dañosos. Por lo tanto, debería considerarse la incriminación, como delitos preparatorios autónomos, del diseño, programación, producción, distribución, venta y compra ilegales de algoritmos, software y sistemas 'maliciosos' de IA. Esta política penal debería limitarse a los sistemas de IA que planteen riesgos elevados para determinados bienes jurídicos significativos (como la vida, la integridad o la libertad de otros seres humanos) y sólo en caso de peligro claro, real y presente (sobre las condiciones exigidas en la incriminación de actos preparatorios, véase la resolución de la Sección I del XVIII Congreso de la AIDP celebrado en Estambul en 2009).

b. Sistemas de IA utilizados en actividades lícitas**:** Este es el caso que suscita las cuestiones más delicadas en referencia al ámbito del 'riesgo permitido', que debería delimitarse mediante la esperada regulación de obligaciones específicas de seguridad y normas cautelares a aplicar a las actividades de diseño, desarrollo, producción, distribución, venta, así como uso, de los sistemas de IA. El ajuste de los modelos de responsabilidad penal en este ámbito debe abordar la fricción que puede crearse entre las formas de responsabilidad por comportamiento negligente y las características técnicas de los sistemas de IA, a saber: (1) su autonomía; (2) la imprevisibilidad concreta de sus decisiones y funcionamiento; (3) la opacidad de sus mecanismos de regulación; (4) la complejidad de su proceso de programación, desarrollo, producción, actualización y mantenimiento.

IV. Sobre la adaptación de los modelos de atribución de responsabilidad a las características de los sistemas de Inteligencia Artificial, en concreto a su grado de autonomía

13. En primer lugar, también en función de la ya reconocida automatización y autonomía graduada de las aplicaciones de la IA en diversos ámbitos, debe distinguirse entre los distintos niveles de autonomía decisoria y operativa de los sistemas de IA, que van desde aquellos en los que el funcionamiento 'automático' permite al agente humano tener un control significativo sobre el sistema, hasta los verdaderamente 'autónomos', en los que la intervención humana se distancia, en el tiempo y en el espacio, del funcionamiento del sistema de IA, que 'decide' en función de la información recogida y de algoritmos que se

adaptan a su experiencia, de modo que existe un margen estructural de imprevisibilidad de los resultados concretos.

14. En relación con los distintos tipos de sistemas de IA, es de fundamental importancia la definición de reglas y normas de funcionamiento específicas, tal y como se prefigura en la propuesta de reglamento europeo sobre Inteligencia Artificial (cfr. *supra* apartado II).

15. La necesidad más acuciante de adaptación de las categorías tradicionales del Derecho Penal se refiere al ámbito de los sistemas de IA con mayor grado de autonomía, que son también el resultado al que tiende el actual desarrollo tecnológico y la experimentación en muchos campos, por lo que sin duda serán aún más importantes en un futuro próximo.

16. Desde esta perspectiva, podrían ser referencia útiles lo desarrollado tanto en el campo de la responsabilidad penal de las personas jurídicas, como en cuanto a la responsabilidad por productos defectuosos y en relación con la protección de la salud y la seguridad en el lugar de trabajo.

17. En estos ámbitos regulados jurídicamente, a menudo armonizados a escala europea, los principios existentes podrían extenderse, con las adaptaciones necesarias, a la regulación de los delitos relacionados con la IA. Dichos ámbitos regulados requieren la evaluación preventiva de los riesgos inherentes a las actividades concretas que se realizan, que tienen márgenes de riesgo permitido y obligaciones correlativas de prevención y contención de riesgos (cfr. apartado II), con especial atención a las fuentes de peligros y daños.

18. Las obligaciones de actuar, especialmente en caso de alertas (*red flags*), se imponen a las categorías pertinentes de personas (seres humanos), que actúan en función de sus competencias respectivas: es decir, los usuarios y las personas en posición de garantía. Deben adaptar con rapidez las medidas reglamentarias y de seguridad de su actividad, hasta el punto de interrumpirla, si fuera necesario.

19. A partir de estos principios reconocidos, pueden elaborarse las siguientes recomendaciones para estructurar la responsabilidad penal por daños relacionados con la IA:

i. Responsabilidad penal de las personas físicas. Debe basarse en la identificación de posiciones personales de garantía, en relación con las competencias y funciones desempeñadas en el uso de sistemas de IA. En primer lugar, habrá que tener en cuenta a los actores que participan en las distintas fases de utilización de los sistemas de IA, hasta llegar a los usuarios finales. En segundo lugar, habrá que considerar las posiciones de los mandos superiores e intermedios y de los responsables del cumplimiento en las organizaciones complejas. En cada caso, se abordará la formalización de obligaciones positivas, de carácter técnico, organizativo y de control.

La responsabilidad penal por conductas negligentes debe ajustarse a los principios generales del Derecho Penal, en concreto, al principio de culpabilidad personal, ya que no basta la conexión objetiva entre la contribución causal del agente humano y la comisión del delito por el sistema de IA, dado que también son necesarias la previsibilidad y evitabilidad del hecho ilícito. En particular, la responsabilidad penal por negligencia, por no haber actuado de forma distinta a la que hubiera sido posible, debe correlacionarse no tanto con el suceso o hecho específico y concreto ocurrido, como con el esquema del 'defecto de organización', atendiendo a la forma en que se estructura y opera el agente artificial. La evaluación de los riesgos derivados de las actividades del sistema de IA también debe incluir la conciencia de que, en casos concretos, es posible que se produzcan resultados impredecibles: conciencia, que constituye la base de la obligación de preparar medidas de vigilancia y contención adecuadas y siempre actualizadas, de las que la persona física encargada sigue siendo responsable, rindiendo cuentas (*accountability*), como propietario o máximo representante de la organización que utiliza el sistema de IA en su propio interés o beneficio.

ii. Responsabilidad punitiva de las personas jurídicas. Teniendo en cuenta que una gran parte de los sistemas de IA son producidos o utilizados por personas jurídicas, es necesario responsabilizarlas de los delitos cometidos por, a través o contra dichos sistemas.

A este respecto, suponiendo que se introduzcan reglas públicas precisas de conducta y cumplimiento (cfr. apartado II), el castigo de la persona jurídica, proporcional a los delitos cometidos por, a través o contra los sistemas de IA y al grado de culpa de la organización, podría basarse en un modelo de responsabilidad basado en el defecto de organización. Dicho modelo de responsabilidad lleva a imputar subjetivamente la responsabilidad, como reprochabilidad culpable, a la persona jurídica en caso de delitos causados por la falta, deficiencia o inadecuación de las medidas organizativas y de prevención, que deben aplicarse y actualizarse con base en la evaluación de los riesgos específicos derivados de las actividades encomendadas y, en todo caso, realizadas por los sistemas de IA, en su interés o en su beneficio. Deberían evitarse los modelos de responsabilidad objetiva (*strict liability*).

Debería promoverse un nuevo modelo de responsabilidad punitiva autónoma de la empresa, que no se base en la responsabilidad de la persona física individual, y que permita considerar a la persona jurídica responsable incluso si la persona física que realizó el daño no es punible individualmente debido a condiciones o circunstancias particulares o si no resulta identificada de manera específica, bastando con constatar la comisión de un acto objetivamente típico e ilícito en interés o en provecho de la organización.

En aquellos ordenamientos jurídicos nacionales que basan la responsabilidad de las empresas en una lista cerrada de delitos, es de recomendar la ampliación de dicha lista a las infracciones penales que puedan cometerse a través, por o contra sistemas de IA.

V. Sobre medidas preventivas y sanciones punitivas aplicables a las personas físicas y jurídicas que se encuentran 'detrás' de los sistemas de Inteligencia Artificial

20. Las sanciones aplicables a las personas físicas, incluidas las penas privativas de libertad, y a las personas jurídicas, eventualmente de carácter administrativo, según los distintos ordenamientos jurídicos, pero en todo caso de carácter punitivo, incluidas las multas y la suspensión de la actividad por la que se cometió la infracción, deberían en principio corresponder a las aplicadas para el tipo de infracción realizada, de conformidad con los principios de cada ordenamiento jurídico particular, a saber, los principios de proporcionalidad y de individualización de las sanciones. Cuando se trate de personas jurídicas, estas sanciones, además de las medidas pecuniarias, deberían incluir también la conminación a modificar el sistema de cumplimiento y control interno de la corporación, así como la posibilidad de ordenar un periodo de vigilancia pública de la misma para asegurar que cumple con las reglas impuestas.

21. Dada la gravedad de los daños relacionados con la IA, es de recomendar la adopción de medidas preventivas por su eficacia para evitar o mitigar las consecuencias perjudiciales de los sistemas de IA (como, entre otras, la incautación, el decomiso, la vigilancia judicial, las medidas interdictivas).

22. Debe subrayarse el importante papel de las penas no pecuniarias, como la sanción de inhabilitación para el ejercicio de actividades específicas y el decomiso. En concreto, el decomiso permite actuar directamente contra el sistema de IA con el que o por el que se cometió el delito, sin necesidad de reconocerle personalidad jurídica o capacidad penal (cfr. apartado I *supra*).

VI. Sistemas coercitivos complementarios

23. Dada la problemática y previsible dificultad de implantar un sistema de responsabilidad penal efectivo para las personas físicas y jurídicas que se encuentran 'detrás' de los sistemas de IA por los delitos cometidos por, a través o contra ellos, se recomienda la adopción de sistemas de ejecución complementarios.

24. Dicho sistema de ejecución podría incluir autorizaciones y certificaciones administrativas, así como recursos civiles.

25. Otras alternativas al enjuiciamiento y/o sanción penal podrían incluir modelos de cumplimiento, intervenciones de justicia restaurativa y acuerdos con las víctimas y las autoridades públicas competentes.

Sección II: La criminalización de los delitos relacionados con la IA

Preámbulo

Teniendo en cuenta que

- La Inteligencia Artificial, uno de los últimos avances de la revolución digital, ha alcanzado ya un importante nivel de desarrollo en esta tercera década del siglo XXI y se utiliza ya ampliamente en muchos sectores de la sociedad;

- Aunque no existe un acuerdo absoluto sobre la definición de esta tecnología, hay un consenso implícito en que incluye multitud de sistemas informatizados que, mediante la recogida, el tratamiento y el análisis de datos en su contexto, son capaces de actuar de forma autónoma y/o ayudar en la toma de decisiones para alcanzar objetivos específicos;

- El potencial transformador de esta tecnología está repercutiendo en múltiples ámbitos y esferas sociales, aportando importantes beneficios y oportunidades;

- La Inteligencia Artificial también plantea riesgos y perjuicios para los intereses individuales y colectivos;

Observando que

- Múltiples decisiones tradicionalmente adoptadas o informadas por humanos empiezan a automatizarse mediante el uso de estas tecnologías, afectando a diferentes ámbitos e intereses;

- En ámbitos como los vehículos autónomos, los servicios sanitarios, los mercados financieros, los medios de comunicación y otros sectores, el uso de esta tecnología es imparable y un futuro sin ella parece inevitable;

- La promesa de que la IA sea eficiente y objetiva está llevando al desarrollo de estas tecnologías sin evaluar su necesidad real ni los riesgos que puede crear;

Teniendo en cuenta que

- los recientes avances en grandes modelos de lenguaje y otros sistemas de IA, como el aprendizaje automático y el aprendizaje profundo, han puesto de manifiesto la necesidad de una regulación, incluidos protocolos de seguridad, para controlar la evolución de estas tecnologías en cuanto a sus efectos y riesgos;

- el desarrollo de sistemas de IA, en particular el entrenamiento de sus algoritmos, requiere el uso y la acumulación de datos y grandes cantidades de información, lo que constituye un riesgo que, en sí mismo, debe considerarse;

Destacando

- la creciente preocupación por el daño que pueden causar los usos malintencionados o negligentes de la IA en ámbitos en los que esta ya empieza a tener una fuerte presencia;

- que hay muchos Estados en los que el uso de sistemas de IA ha causado daños a intereses relevantes como la vida, la salud y la privacidad, entre otros;

Reconociendo

- que la aparición de nuevos actos delictivos, así como de nuevos intereses dignos de protección penal, llevará a los Estados a adaptar las leyes penales relacionadas con la IA;

- la necesidad de analizar si la respuesta legal de los Estados a los retos de la IA es suficiente o necesita ser reformada y adaptada, ya sea mediante modificaciones específicas o mediante la creación de nuevos tipos penales;

Teniendo en cuenta

- la Recomendación del Consejo sobre Inteligencia Artificial adoptada por la OCD el 22 de mayo de 2019; la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre Inteligencia Artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión presentada por la Comisión Europea el 21 de abril de 2021; la Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su uso por las autoridades policiales y judiciales en materia penal (2020/2016(INI)); las Directrices éticas para una IA digna de confianza presentadas por el Grupo de Expertos de Alto Nivel creado por la Comisión Europea el 8 de abril de 2019; y la Recomendación CM/Rec (2020) 1 del Comité de Ministros a los Estados miembros sobre el impacto de los sistemas algorítmicos en los derechos humanos (Adoptada por el Comité de Ministros el 8 de abril de 2020 en la 1373ª reunión de los Delegados de los Ministros);

- los debates y resoluciones de anteriores Congresos Internacionales de la Asociación Internacional de Derecho Penal, en particular las resoluciones del XIV Congreso sobre Derecho penal y técnicas biomédicas modernas y los problemas jurídicos y prácticos que plantea la diferencia entre Derecho penal y Derecho administrativo sancionador; las resoluciones del XIX Congreso Internacional de Derecho Penal - Sociedad de la información y Derecho penal; y los proyectos de resolución del XXI Congreso Internacional de Derecho Penal sobre la Sección I (Derecho penal - parte general) y la Sección 3 (La inteligencia artificial y la administración de justicia penal);

Los participantes en el Coloquio Internacional de la Sección II, celebrado en Bucarest, del 14 al 16 de junio de 2023, han aprobado lo siguiente:

Resoluciones

I. Regulación e IA

1. Con el desarrollo y la expansión de la IA han surgido y es probable que aumenten los daños y riesgos para los intereses individuales y colectivos. Esto plantea la necesidad de una regulación adecuada de la producción de los sistemas de IA y de su uso. La respuesta a estos retos deberá considerar diferentes perspectivas y todas las herramientas legales disponibles, ya sean públicas o privadas, teniendo en cuenta la diferente naturaleza y funcionalidad de cada una de ellas.

2. El impacto global de la IA exige una respuesta internacional para proteger eficazmente los intereses individuales y colectivos en juego. Los Estados deben tener en cuenta las normas internacionales en la regulación y ejecución de las normas nacionales.

3. Como ha sucedido con otras innovaciones tecnológicas y socioeconómicas, por ejemplo, Internet o los nuevos desarrollos en edición genética y neurociencia, la irrupción de la IA hace necesario revisar aspectos generales del sistema de justicia penal y, en particular, el catálogo de delitos existentes en las leyes penales para comprobar si la regulación de los Estados se adecua a los retos que plantea el uso de esta tecnología.

4. Los debates sobre la transformación del Derecho penal en respuesta al impacto de la IA, y el papel que el sistema de Derecho penal puede desempeñar en relación con estas nuevas tecnologías, no deben aislarse de los debates éticos internacionales sobre el desarrollo de la IA, ni pueden pasar por alto los debates sobre la regulación de la IA en otras ramas del ordenamiento jurídico. Teniendo en cuenta este amplio panorama, el legislador debe reflexionar sobre el papel específico que desempeña el Derecho penal para evitar los daños causados por el uso de la IA.

5. Los legisladores tendrán que reformar los delitos existentes cuando la IA modifique el riesgo para los intereses existentes o cree nuevos medios de perpetración que no estén cubiertos por la legislación vigente. Además, se introducirán nuevos delitos cuando el desarrollo de esta tecnología dé lugar a la aparición de nuevos intereses individuales y colectivos dignos de protección que no estén cubiertos por la legislación vigente.

II. Criminalización y protección de los intereses relacionados con la IA

6. El desarrollo de la IA puede suponer la aparición de nuevos intereses dignos de tutela. Además, los sistemas de IA pueden afectar a la dimensión y relevancia de intereses que actualmente no se consideran dignos de protección por el Derecho penal. Cuando las leyes penales no ofrezcan una respuesta adecuada para proteger estos intereses, se

promulgarán nuevos tipos penales que castiguen proporcionalmente las conductas lesivas para dichos intereses. Esto solo se hará cuando no existan medios menos lesivos que el Derecho penal para proteger eficazmente tales intereses.

7. Cuando la transformación de la IA da lugar a la aparición de nuevos intereses que son esencialmente similares a otros que tradicionalmente se consideraban dignos de protección no deben introducirse nuevas figuras delictivas.

En su lugar, es preferible adaptar la interpretación de los delitos existentes, siempre que lo permita el estricto respeto del principio de legalidad.

8. Los legisladores tendrán que decidir si el desarrollo de esta tecnología da lugar a la necesidad de una protección penal específica de los intereses individuales o colectivos relacionados con la propia tecnología de IA. Aunque todavía es demasiado pronto para determinar si surgirá tal necesidad, podría ser el caso de los datos en los que se basan los algoritmos; la funcionalidad de los propios sistemas de IA, en algunos casos; los intereses colectivos relacionados con la seguridad y fiabilidad de su diseño y aplicación; o incluso los intereses asociados a los robots.

9. Algunos sistemas de IA, como los utilizados en infraestructuras críticas, son esenciales para garantizar intereses ya protegidos. En la medida en que el Convenio de Budapest contempla la penalización de los ataques a los sistemas informáticos y la IA puede considerarse como tal, la promulgación de nuevos delitos podría no ser necesaria. Para no suscitar dudas interpretativas, podría ser aconsejable reformar algunos delitos para introducir los sistemas de IA como un tipo de sistema informático.

10. Si no se demuestra que hay otros intereses en juego y teniendo en cuenta el nivel actual de desarrollo de estas tecnologías, la IA y los sistemas robóticos no requieren de una protección penal diferente, en relación con su valor económico o funcional, de la que ya disponen otros sistemas informáticos.

III. Bases de legitimación y técnicas de criminalización

11. El Derecho penal no debe desempeñar un papel protagonista en la regulación de la IA. Dada su naturaleza de instrumento coercitivo especialmente lesivo, debe intervenir como último recurso y limitarse a la represión de los actos más graves y perjudiciales.

12. Los legisladores no deben introducir nuevos delitos basados únicamente en el hecho de que se haya empleado IA en su comisión. Muchos delitos pueden cometerse utilizando sistemas de IA como medio para llevar a cabo la conducta sancionada. Solo cuando los actos cometidos con sistemas de IA adquieran un significado diferente en términos de lesividad o riesgo será necesario introducir nuevas figuras delictivas.

13. La automatización de los procesos de toma de decisiones basados en datos que conlleva la IA resitúa el momento clave de la agencia humana en fases de diseño e implementación de algoritmos muy alejadas del daño. Por lo tanto, la responsabilidad de las personas físicas y jurídicas implicadas debería centrarse preferentemente en estas fases. Esto se hará teniendo en cuenta las obligaciones existentes establecidos en otras ramas del ordenamiento jurídico.

14. Los sistemas penales están diseñados para tener un efecto disuasorio sobre los infractores, previniendo que cometan acciones delictivas. Si el momento clave en términos de riesgo en relación con la IA es el momento de su diseño y puesta en práctica, deberá considerarse la promulgación de delitos que tengan por objeto disuadir conductas en esos momentos. Esto puede hacerse anticipando la protección con delitos de peligro que castiguen el incumplimiento de determinados deberes de conducta en relación con intereses concretos dignos de protección. Asimismo, y de forma similar a lo predicado respecto al régimen de responsabilidad de las personas jurídicas, podrían establecerse obligaciones específicas relacionadas con el diseño y la implantación de sistemas de IA, cuya infracción podría dar lugar a responsabilidad penal.

15. Los delitos de peligro relacionados con el diseño y la implementación de sistemas de IA que causen riesgos se promulgarán cuando las acciones sancionadas supongan una amenaza considerable para los intereses protegidos. Además, las consecuencias jurídicas de estas infracciones serán proporcionales al nivel de riesgo causado y al interés en juego. Debido a la complejidad del diseño de los sistemas de IA y a los diferentes enfoques de regulación de estas herramientas, no se tipificarán los delitos de peligro antes de considerar la evolución de la autorregulación o de la normativa administrativa sobre control y seguridad de la IA en cada ordenamiento jurídico. Esta normativa servirá para identificar los actos de riesgo relevantes que puedan ser merecedores de persecución penal.

16. Podrían introducirse nuevos delitos para castigar el abuso y la transformación de los sistemas de IA legales existentes cuando, al cambiar el diseño o la finalidad de la IA, surjan nuevos riesgos.

17. En los ordenamientos jurídicos en los que la negligencia sólo se castiga cuando está expresamente prevista (*numerus clausus*), podría ser necesaria una reforma de la legislación penal. El complejo diseño de la IA y la participación de múltiples partes en el ciclo de vida de la IA implica que en la mayoría de los casos será extremadamente difícil demostrar que en el momento del diseño los sujetos se representasen la producción de un resultado perjudicial. Los delitos imprudentes basados en la infracción de las normas de cuidado podrían entonces promulgarse si la protección de los intereses afectados lo hiciera necesario.

18. Dado que los sistemas de IA son dinámicos y su funcionamiento depende de la introducción o recopilación de datos que modifican sus resultados, los procesos de gestión de riesgos establecidos en otras ramas del ordenamiento jurídico podrían operar

a lo largo de todo el ciclo de vida de la IA. Así, las leyes penales pueden abordar, en caso necesario, las infracciones de las normas relacionadas con la falta de seguimiento y supervisión adecuados de los sistemas de IA, obligaciones que podrían afectar a diferentes sujetos implicados en todo el ciclo de vida de la IA.

19. Dado que la tecnología de IA y sus aplicaciones pueden hacer escalar o multiplicar la incidencia de ciertos procesos, los sistemas de justicia penal podrían adaptarse para ajustar la proporcionalidad de las penas a la gravedad del daño que puede causar la IA. No obstante, los legisladores no deben promulgar circunstancias agravantes solo porque un delito se haya cometido utilizando IA. Solo en el caso de que las circunstancias agravantes existentes no puedan contemplar la gravedad de los daños causados por el uso de la IA, atendiendo también a la relevancia de los intereses afectados, se considerarán nuevas formas de agravación. Esto se hará siempre respetando el principio de proporcionalidad.

IV. Criminalización y protección de intereses específicos frente a los riesgos creados por la IA.

20. Dado que las leyes penales no suelen prever medios específicos para la comisión de delitos contra la vida y la salud, no parece necesario reformar estos delitos para proteger tales intereses cuando se ha utilizado el sistema de IA como medio comisivo. Tampoco parece necesario modificar el sistema de graduación de la responsabilidad. No obstante, en áreas específicas, como la conducción autónoma, el Derecho penal tendrá que considerar el cambiante panorama normativo, que será la base para afirmar lo que se considera un riesgo permitido para la determinación de la responsabilidad.

21. Si la conducción autónoma se generaliza, los delitos contra la seguridad vial podrían sufrir cambios significativos, incluyendo nuevos delitos relacionados con nuevas conductas de riesgo para la vida y la seguridad vial distintas de las que actualmente se centran en la conducción humana.

22. De la misma manera que la revolución sobre la edición genética llevó a la aparición de delitos que sancionan la manipulación genética con capacidad de destrucción masiva, la evolución tecnológica puede hacer necesario, en un futuro próximo, criminalizar la creación, desarrollo y uso de herramientas de IA con alta capacidad destructiva, como algunas armas autónomas, drones o robots que podrían ser enormemente dañinos especialmente si se pierde el control humano.

23. Los sistemas de IA recopilan y utilizan grandes cantidades de información para realizar sus tareas, lo que crea nuevas amenazas para intereses que clásicamente han sido objeto de protección por el Derecho penal. Dado este desarrollo de la tecnología de IA, es necesaria una revisión de los delitos relacionados con la intimidad y otros intereses personales, y debería tenerse en cuenta una revisión de la concepción de la intimidad como un bien exclusivamente individual, considerando la dimensión colectiva de este interés.

24. La criminalización de los actos que implican la recopilación ilícita de datos personales no debe vincularse únicamente a la protección de intereses como la privacidad. El uso de la IA en el ciberespacio puede abrir la puerta a la recogida masiva de datos para la comisión de ciberdelitos que lesionen intereses como la propiedad. Los Estados deben revisar si es necesario criminalizar conductas como la recopilación masiva ilegal de datos, y actos preparatorios respecto a delitos graves, en los que casos que supongan un riesgo concreto para esos intereses y solo si no existen otros instrumentos menos invasivos que el Derecho penal para dar una respuesta adecuada.

25. La accesibilidad de imágenes y datos personales en el ciberespacio unida al potencial de la IA generativa para transformar imágenes, vídeo y audio puede poner en peligro intereses como el honor o la libertad sexual. Es necesario revisar si las leyes penales actuales permiten castigar conductas lesivas para la dignidad humana, el honor y la libertad sexual como la distribución de *Deep Fakes*, incluidos los de contenido sexual, o de pornografía infantil.

26. Las IA generativas, como los grandes modelos de lenguaje y herramientas similares, pueden facilitar el engaño, la amenaza y la coacción, afectando a distintas fases de la formación de la voluntad, poniendo en peligro intereses dignos de protección. Sin embargo, no parece adecuado introducir nuevos tipos penales, ya que las leyes penales actuales abarcan los actos más dañinos y deberían utilizarse otros medios de control de este tipo de herramientas para prevenir conductas de menor riesgo.

27. La popularización de algoritmos para la gestión de riesgos en ámbitos como la sanidad, la contratación de empleados, la justicia, la concesión de créditos y préstamos, y muchos otros, ha revelado la existencia de sesgos discriminatorios en algunas decisiones tomadas por los sistemas de IA. Más allá de los delitos que ya pueden castigar algunas decisiones discriminatorias especialmente graves, otras ramas del ordenamiento jurídico, como el Derecho civil o administrativo, son más apropiadas para evitar el problema de la discriminación algorítmica.

28. El uso de la IA en el ciberespacio puede facilitar y potenciar los ataques existentes contra la propiedad y otros intereses dignos de protección. Sin embargo, dada la regulación del ciberfraude y otros ciberdelitos contra la propiedad, no será necesario, al menos a corto plazo, adaptar estos delitos para dar cabida a los comportamientos perpetrados mediante sistemas de IA.

29. Algunas de las leyes penales que castigan la producción, la venta, la adquisición para su uso, la importación, la distribución o cualquier otra forma de puesta a disposición de dispositivos diseñados o adaptados principalmente para cometer delitos, promulgadas de conformidad con los artículos 2 a 5 del Convenio de Budapest, ya pueden castigar la creación, el desarrollo y la venta de sistemas de IA diseñados o adaptados para esos fines delictivos. Por lo tanto, si la IA se considera un dispositivo, incluido un programa informático, de conformidad con el Art. 6 del Convenio de Budapest, la introducción de nuevos delitos que anticipen la respuesta penal no es necesaria en este ámbito.

30. En el ámbito socioeconómico y financiero, ya se está denunciando la proliferación de sistemas algorítmicos de toma de decisiones y el uso de la IA para el comercio. El riesgo de que el uso malintencionado o negligente de los sistemas de IA afecte gravemente a los mercados es evidente, pero los intereses en juego podrían protegerse mejor con medidas regulatorias preventivas de carácter económico, administrativo y mercantil, que con sanciones penales distintas a las ya existentes para castigar el uso de información privilegiada y conductas similares. Cuando los sistemas de IA se utilizan para manipular los mercados, el Derecho penal debería revisarse para dar una respuesta proporcional.

31. La preocupación por el impacto de la desinformación, primero tras algunos procesos electorales y después con la infodemia durante la crisis del COVID-19, ha llevado a muchos Estados a crear figuras penales para castigar esta conducta. El hecho de que la IA pueda aumentar su impacto, ya sea automatizando su difusión o utilizando sofisticadas tecnologías de manipulación de vídeo, audio, imagen y texto, puede mantener esta tendencia. La criminalización de la desinformación sólo estará justificada para la protección de los intereses fundamentales de las sociedades democráticas y si no pone en peligro la libertad de expresión.

Sección III: La IA y la administración de justicia penal: 'policía predictiva', 'justicia predictiva' y derecho probatorio

Preámbulo

Conscientes de que la inteligencia artificial (IA) se está desarrollando rápidamente en la sociedad contemporánea en diversas regiones del mundo. En algunos países ya es omnipresente en la vida de las personas, y en el futuro podría formar parte de la vida cotidiana de un gran sector de la población mundial.

Reconociendo que, como innovación tecnológica, la IA impulsa a los consumidores a adquirir nuevos productos, contribuyendo así al crecimiento de la economía global. Por lo tanto, la IA desempeña un papel no desdeñable en el sostenimiento e, incluso, en la expansión de la economía liberal de mercado y del sistema económico capitalista.

Reconociendo que las empresas que crean y comercializan la IA suelen tener su sede en países desarrollados del hemisferio norte y, a menudo, intentan abrir mercados en todo el mundo.

Reconociendo que la 'brecha digital' amplía las desigualdades sociales entre las personas. La 'brecha de la IA' podría ser el próximo gran desafío.

***

Considerando que la IA puede definirse como un conjunto de teorías y técnicas utilizadas para crear máquinas capaces de simular la inteligencia humana[1]. Como

[1] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

disciplina científica, combina matemáticas estadísticas y algorítmicas, informática y ciencias cognitivas. La IA simbólica se basa en las reglas de la lógica, mientras que la IA conexionista utiliza redes neuronales artificiales.

Considerando que el aprendizaje automático es un ejemplo de IA conexionista, al igual que el aprendizaje profundo, que es una subcategoría del aprendizaje automático que emplea múltiples capas de neuronas artificiales interconectadas. A medida que aumenta el número de capas neuronales que posibilitan el aprendizaje autónomo, también aumenta la complejidad tecnológica del sistema, lo que hace que el sistema sea más eficiente y sus cálculos menos explicables y rastreables (aprendizaje profundo).

Considerando que el aprendizaje automático puede hacer que la tecnología sea extremadamente potente, aunque su proceso de toma de decisiones puede ser tan complejo que se asemeje a una 'caja negra'.

Reconociendo que muchos sistemas de IA utilizados en el contexto de la prevención, investigación, detección y castigo de delitos son sistemas de aprendizaje automático. Utilizando algoritmos de autoaprendizaje, realizan complejos cálculos de probabilidades en nanosegundos. Para lograr los objetivos que se les asignan, procesan enormes cantidades de datos y consumen mucha energía. Algunos de ellos, como los sistemas de reconocimiento facial, se basan en el aprendizaje profundo.

***

Observando que, en la administración de justicia penal, los sistemas de IA se utilizan para prevenir o detectar delitos mediante la evaluación de riesgos ('policía predictiva').

Observando que los sistemas de IA también se emplean para asistir a los fiscales y jueces en la toma de decisiones. En particular, el término 'justicia predictiva' hace referencia a (i) la anticipación del comportamiento de una persona, por ejemplo, para evaluar el riesgo de fuga en el procedimiento previo al juicio o de volver a cometer un delito, de modo que se puedan tomar decisiones sobre ella, como la prisión preventiva, la condena, la libertad condicional y la libertad sometida a vigilancia (*probation*) (justicia actuarial, que hoy en día puede estar respaldada por la IA); y (ii) el uso de la IA para realizar análisis estadísticos ultrarrápidos de decisiones previas dictadas en casos similares y de las disposiciones legales y reglamentarias relevantes (análisis jurídico cuantitativo o LegalTech).

Considerando que el análisis jurídico cuantitativo es revolucionario, en el sentido de que un cálculo matemático está destinado a apoyar o incluso a suplantar el razonamiento jurídico.

Observando que el término 'predictivo' utilizado en expresiones como 'policía predictiva' o 'justicia predictiva' es confuso, porque los sistemas de IA calculan probabilidades, pero no predicen el futuro; dichas probabilidades se basan en correlaciones, no en causalidades. Estos cálculos, sin embargo, tienen un efecto performativo en las personas, es decir, pueden inducirlas a tomar decisiones alineadas

con los resultados obtenidos. En términos generales, las bases científicas de la IA fomentan que los usuarios confíen y sigan las probabilidades calculadas por el sistema de IA, ya que el 'sesgo de automatización' aumenta cuando el sistema se percibe con un cierto grado de aura científica.

Observando que los sistemas de IA contribuyen a la innovación en la búsqueda de pruebas. Son capaces de analizar con rapidez grandes cantidades de datos y extraer información útil para los investigadores. Los sistemas de IA pueden establecer correlaciones entre datos que son imperceptibles para el ojo humano. Los diagramas de análisis de la delincuencia que generan pueden proporcionar información altamente valiosa y elaborada para los investigadores.

Observando que los sistemas de IA pueden generar información presentada como prueba para su uso en juicios penales. En particular, los sistemas de IA pueden proporcionar información forense comparando rasgos biométricos (por ejemplo, imágenes faciales en el reconocimiento facial), frecuencias de sonido de diferentes voces (reconocimiento vocal) y fragmentos de ADN (genotipado probabilístico).

Observando, por último, que los robots asistidos por IA y los objetos 'inteligentes' en diversos ámbitos laborales y de la vida cotidiana pueden, de manera incidental, generar indicios o pruebas que podrían resultar útiles para establecer los hechos en un caso penal.

***

Observando que, a pesar de los importantes avances de los últimos años, los sistemas de IA no son completamente fiables. Los errores pueden deberse a la mala calidad de los datos utilizados, a la forma en que se programa el algoritmo o a la existencia de falsos positivos/negativos en las correlaciones. Por lo tanto, las probabilidades generadas por un sistema de IA pueden ser inexactas.

Observando que los resultados producidos por los sistemas de IA no siempre son completamente neutrales. La precisión de las probabilidades calculadas por los sistemas de IA depende no sólo de la calidad de los datos recopilados y procesados, los cuales pueden reflejar sesgos, sino también de la forma en que estos sistemas han aprendido (aprendizaje no supervisado frente a aprendizaje supervisado). Como reproducen decisiones humanas, los algoritmos de autoaprendizaje se ven influidos por las debilidades humanas. Un resultado son los algoritmos xenófobos[2], racistas, misóginos, etc.

Observando que los sistemas de IA presentan problemas de transparencia. La llamada 'caja negra de la IA' es tan opaca que ni siquiera los especialistas pueden determinar cómo llega a sus resultados. Ni siquiera los expertos científicos pueden explicar completamente el razonamiento de un sistema ante un tribunal.

[2] https://www.amnesty.org/en/documents/eur35/4686/2021/en/

Considerando que los sistemas de IA utilizados en el ámbito de la justicia penal pueden ser desarrollados por el sector privado. Dichos sistemas son productos destinados a la venta y deben ser rentables. Las empresas que los desarrollan invocan generalmente la protección del secreto comercial para negarse a revelar el código fuente de sus algoritmos, sin el cual no se puede analizar adecuadamente el funcionamiento del sistema.

***

Observando que no todo lo que es tecnológicamente posible es socialmente deseable. En una democracia, las decisiones políticas que afectan a la prevención, detección, investigación y sanción de los delitos deben plasmarse en una ley o en una norma de fuerza vinculante equivalente.

Reiterando que los derechos humanos deben ser plenamente protegidos al prevenir, detectar, investigar y sancionar delitos, incluso cuando se utilicen innovaciones tecnológicas en ese contexto. Dado que la IA plantea a menudo cuestiones relativas a la privacidad y la protección de datos personales, así como al derecho a la no discriminación, todas las leyes que protegen a los seres humanos, en particular su libertad y dignidad, así como todas las garantías de un juicio justo, incluida la presunción de inocencia, se ven potencialmente amenazadas por el uso de tecnologías que simulan la inteligencia humana.

Considerando que las leyes nacionales y las normas jurídicas internacionales y/o regionales pueden establecer las condiciones bajo las cuales se puede permitir que las innovaciones tecnológicas relacionadas con la IA contribuyan a la administración de justicia penal.

Reiterando que los estándares éticos a los que frecuentemente se refiere el sector privado no tienen la misma fuerza vinculante que la ley.

***

Tomando en cuenta:

* La Recomendación del Consejo sobre Inteligencia Artificial, Organización para la Cooperación y el Desarrollo Económicos, 22 de mayo de 2019, C/MIN(2019)3/FINAL;

* La Recomendación sobre la ética de la inteligencia artificial, Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 23 de noviembre de 2021, SHS/BIO/PI/2021/1;

* La Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno, Comisión Europea para la Eficiencia de la Justicia, Estrasburgo, 3-4 de diciembre de 2018;

* La Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales, documento 2020, 2016 (INI);

* Las resoluciones del XIX Congreso Internacional de Derecho Penal: Sociedad de la información y derecho penal, Río de Janeiro, 2014.

***

Los participantes en el Coloquio Internacional de la Sección III, celebrado en Buenos Aires, del 28 al 31 de marzo de 2023, han aprobado lo siguiente:

Resoluciones

1. El empleo de sistemas de IA por parte de las autoridades públicas para asistir en la prevención, detección o investigación de delitos debe estar autorizado previamente por una ley o una norma con fuerza vinculante equivalente[3].

2. Los Estados deben garantizar que las decisiones adoptadas por las autoridades para centrarse en la prevención, detección o investigación de un tipo particular de delito estén basadas en criterios determinados política y democráticamente, y no en la suposición de que el uso de tecnología de IA facilitará dicha tares.

3. Para proteger la legitimidad de las actividades de las autoridades públicas en la prevención, detección e investigación de delitos, los Estados que deseen utilizar sistemas de IA deben elegir aquellos sistemas cuyo funcionamiento sea totalmente transparente, explicable y rastreable (modelo de IA de caja blanca). Deben garantizar que las objeciones relacionadas con la propiedad intelectual no impidan la búsqueda de transparencia, y deberían optar preferentemente por sistemas de código abierto y accesibles al público.

4. Las leyes o normas equivalentes relativas al uso de sistemas de IA para la prevención, detección e investigación de delitos deberían exigir que dichos sistemas cuenten con un alto grado de fiabilidad tecnológica. Una regulación suficientemente precisa que exija verificaciones y evaluaciones apropiadas, tanto externas como independientes del desarrollador y proveedor del sistema de IA, debería limitar en la mayor medida posible el riesgo de sesgo o cualquier forma de discriminación en el aprendizaje automático, errores de codificación y otros fallos tecnológicos.

5. Las leyes o normas equivalentes deben exigir que los sistemas de IA utilizados para asistir en la prevención, detección e investigación de delitos sean totalmente accesibles, verificables y auditables por las autoridades que los utilicen y por las autoridades encargadas de realizar las verificaciones y evaluaciones.

[3] En lo sucesivo, utilizaremos la expresión 'norma equivalente' para referirnos a la 'norma con fuerza vinculante equivalente'.

6. Las leyes o normas equivalentes que autorizan el uso de sistemas de IA para asistir en la prevención, detección e investigación de delitos deben exigir que los datos de entrenamiento sean de alta calidad y representativos.

En cuanto a los datos procedentes de archivos policiales o judiciales, las leyes o normas equivalentes deben establecer un mecanismo que garantice que dichos datos sean correctos, estén actualizados y que su uso no infrinja la presunción de inocencia. La presunción de inocencia prohíbe terminantemente la conservación y el uso de datos recopilados como consecuencia de una evaluación predictiva cuando no haya un posterior fallo de culpabilidad, excepto si los datos son relevantes en relación con otro sospechoso.

En cuanto a otros datos, en particular los datos accesibles en las redes sociales, las leyes o normas equivalentes deben exigir el cumplimiento del derecho a la privacidad y de la legislación de protección de datos personales al utilizar dichos datos. Deben realizarse verificaciones adecuadas, independientes de las instituciones policiales y judiciales.

En general, las leyes o normas equivalentes deben ser altamente exigentes respecto a la verificación de la fiabilidad de todos los datos utilizados por los sistemas de IA en relación con la detección, prevención e investigación de delitos.

7. Las leyes o normas equivalentes deben exigir que, antes de que un sistema de IA basado en algoritmos de autoaprendizaje pueda ser utilizado para prevenir, detectar o investigar delitos, los algoritmos sean desarrollados, entrenados, probados y desplegados bajo supervisión humana (aprendizaje automático con intervención humana - *human-in-the-loop machine learning*).

Estas leyes y normas equivalentes deben exigir una evaluación humana antes de que se adopte cualquier medida para prevenir, detectar o investigar delitos basándose en las probabilidades calculadas por un sistema de IA.

8. Los Estados y las autoridades encargadas de la aplicación de la ley deben garantizar que su personal, al utilizar la IA para prevenir, detectar o investigar delitos, reciba formación práctica sobre el uso adecuado del sistema de IA correspondiente, así como formación sobre los riesgos asociados a errores y sesgos. Deben garantizar que dicho personal tenga un conocimiento profundo de los peligros que la IA puede representar para los derechos humanos.

9. Las autoridades internacionales, regionales, nacionales o locales deben establecer organismos independientes encargados de certificar la calidad de los sistemas de IA destinados a ser utilizados en la prevención, detección o investigación de delitos. No se debe certificar ninguna tecnología de IA que no pueda ser operada y supervisada de manera transparente, entre otras razones, debido a los derechos de propiedad intelectual.

El sector privado debería organizarse o unirse para crear etiquetas de calidad para los sistemas de IA con el objetivo de fomentar un círculo virtuoso en torno a estos productos,

de modo que las autoridades encargadas de prevenir, detectar o investigar delitos puedan determinar mejor qué sistemas de IA se ajustan a sus necesidades.

10. Cuando se utilizan sistemas de IA para prevenir, detectar o investigar delitos deben protegerse todos los derechos humanos. Los Estados y los organismos regionales e internacionales deben asegurar la imposición de sanciones efectivas, proporcionadas y disuasorias cuando se violen dichos derechos.

Las leyes o normas equivalentes deben establecer expresamente que, cuando la causa de la violación de los derechos humanos sea el mal funcionamiento tecnológico de un sistema de IA, la empresa que desarrolló el sistema sea responsable por culpa, negligencia o en virtud del criterio de la responsabilidad objetiva por productos defectuosos. Asimismo, deben prever que se lleven a cabo investigaciones para determinar la causa de la violación.

11. Las presentes resoluciones son igualmente aplicables a la prevención, detección, investigación y sanción de infracciones administrativas por parte de las autoridades competentes.

Resoluciones específicas sobre policía predictiva

12. Los Estados y los organismos regionales e internacionales de derechos humanos deben garantizar que el uso de sistemas de IA en la prevención y detección de delitos no derive en prácticas de vigilancia masiva, lo que provocaría una reducción desproporcionada de las libertades individuales (libertad de circulación, libertad de expresión, libertad de reunión, libertad de asociación y libertad de religión).

En particular, los Estados y las autoridades locales deben prohibir el uso de sistemas de IA para la identificación remota de personas en espacios de acceso público basada en sus datos biométricos, así como cualquier otra aplicación de sistemas de IA que permita la vigilancia masiva.

Se insta a los Estados a ser más transparentes sobre el uso de sistemas automatizados de reconocimiento de matrículas en lugares públicos. Cuando estos sistemas incluyan no sólo la toma de imágenes de la matrícula, sino también de cualquier persona que se encuentre en el vehículo, esta opción debe estar autorizada explícitamente por la ley. Debe prohibirse la aplicación de la tecnología de reconocimiento facial a los datos obtenidos de estas imágenes con fines de policía predictiva, salvo en el contexto de una investigación específica si existe un marco legal para ello.

13. Los Estados deben determinar o encargar a organismos de investigación independientes que establezcan si el uso de sistemas de IA en la prevención de delitos ayuda a reducir el número de delitos cometidos y, de ser así, en qué proporción.

14. Los Estados deben garantizar que el coste financiero de los sistemas de IA y su mantenimiento no prive de fondos a los servicios públicos de prevención del delito que

trabajan sobre las causas de la delincuencia (para apoyo psicológico, apoyo social, formación y apoyo al empleo).

15. Las leyes y las normas equivalentes deben prohibir de forma estricta el uso de datos como prueba incriminatoria en procedimientos penales cuando dichos datos hayan sido recopilados por un sistema de IA en el contexto de la prevención de la delincuencia, es decir, cuando no existía una sospecha concreta de que se hubiera cometido un delito y, por lo tanto, los datos se hayan recogido fuera del marco legal que regula las investigaciones penales.

Si los datos recogidos por un sistema de IA en el contexto de la prevención del delito se utilizan en la investigación penal como información de partida, la autoridad judicial competente debe ser informada de ello. Los datos deben ser marcados como tales y el uso de sistemas de IA debe quedar documentado en el expediente del caso.

Resoluciones específicas sobre justicia predictiva

16. Las leyes y normas equivalentes deben prohibir de manera estricta el uso de sistemas de IA con fines de justicia actuarial en la determinación de la pena.

Castigar o agravar la pena de una persona basándose en la probabilidad de que cometa un delito en el futuro equivale a imponer un castigo basado, en parte, en un acto delictivo que no ha ocurrido. Esto resulta contrario a la dignidad humana, la libertad personal y los principios fundamentales de la justicia penal.

Debe prohibirse el uso de herramientas de evaluación del riesgo basadas en la IA para justificar medidas de seguridad severas, como la detención. Cuando los Estados permitan su uso para medidas menos severas, la ley debe autorizarlo expresamente y prever suficientes garantías procesales. Sin embargo, las probabilidades generadas por la IA no pueden constituir el único fundamento de una decisión.

17. Los Estados que deseen utilizar la IA para asistir a los fiscales y/o jueces en el análisis jurídico cuantitativo antes de tomar decisiones en causas penales deben limitar su uso a delitos menores que impliquen un elevado volumen de casos.

18. Antes de tomar la decisión de utilizar la IA para facilitar la gestión de un gran volumen de casos relativos a delitos menores, los Estados deben evaluar si, a la luz del principio de *ultima ratio*, sería más adecuado despenalizar las conductas que generan dichos casos.

19. Las leyes y las normas equivalentes deben prohibir el uso del análisis jurídico cuantitativo para asistir a los jueces a la hora de pronunciarse sobre la culpabilidad.

20. Las leyes y normas equivalentes deben prohibir el uso del análisis jurídico cuantitativo para asistir a los jueces en la determinación de las penas. La decisión de castigar a una persona y el tipo de pena debe ser tomada por seres humanos. De lo

contrario, la justicia podría deshumanizarse y la dignidad humana de las personas quedar amenazada.

21. Las leyes y normas equivalentes deben prohibir el uso del análisis jurídico cuantitativo para asistir a los jueces en las decisiones en materia penal que se dicten antes de la sentencia y que impliquen medidas coercitivas.

22. Los Estados deben garantizar que las decisiones adoptadas con la ayuda del análisis jurídico cuantitativo no infrinjan el derecho a acceder a un juez humano.

23. Las leyes y normas equivalentes deben prohibir el uso del análisis jurídico cuantitativo a menos que la decisión pueda ser recurrida por la persona afectada. La decisión en la instancia de apelación no se basará únicamente en el análisis jurídico cuantitativo.

Resoluciones específicas sobre las pruebas obtenidas y/o producidas por la IA

24. Las leyes y normas equivalentes sobre la extracción de datos para su análisis por sistemas de IA deben exigir que, antes de solicitar a una persona el código de acceso a su software o hardware del que se puedan extraer datos, la autoridad que realiza la incautación debe informar a la persona afectada de su derecho a no autoincriminarse.

25. Las leyes y normas equivalentes sobre el análisis de la delincuencia deben especificar que los diagramas de análisis de la delincuencia producidos por los sistemas de IA carecen de valor probatorio, pero pueden servir de guía para la orientar la investigación.

26. Las leyes y normas equivalentes sobre el uso de sistemas de IA para recopilar pruebas o generar información con fines de justicia penal deben indicar claramente que los resultados de dichos sistemas son sólo probabilidades. Deben exigir que cualquier juicio basado en probabilidades indique no sólo la probabilidad calculada por el sistema de IA utilizado, sino también la tasa de error de dicho sistema, según lo calculado por el organismo de certificación que lo evaluó.

27. Los Estados y las autoridades judiciales deben garantizar que el uso de probabilidades calculadas por la IA no disminuya el estándar de prueba existente en los procedimientos penales.

28. Las leyes y normas equivalentes sobre el uso de sistemas de IA para recopilar pruebas o generar información con fines de justicia penal deben prohibir el uso, como prueba, de probabilidades calculadas por sistemas de IA que no sean totalmente explicables (IA de caja negra).

29. Las leyes y normas equivalentes sobre el uso de sistemas de IA para recopilar pruebas o generar información con fines de justicia penal deben exigir, de conformidad con el derecho a la contradicción, que, si se utilizan datos recogidos o producidos por un sistema de IA, todas las partes sean informadas de ello. Los datos deben marcarse como tales y el uso de sistemas de IA debe documentarse en el expediente del caso.

Las leyes y las normas equivalentes deben exigir que la presentación por una de las partes de una probabilidad calculada por la IA pueda ser impugnada por la otra parte.

30. Las leyes y normas equivalentes deben establecer el principio de que la parte que presente la probabilidad ante el tribunal debe proporcionar sistemáticamente información completa sobre el funcionamiento del sistema de IA y los datos que utiliza.

31. Las leyes y normas equivalentes sobre el uso de sistemas de IA para recopilar pruebas o generar información con fines de justicia penal deben garantizar, de conformidad con los derechos de la defensa, que cualquier persona acusada de un delito con base en una probabilidad presentada como prueba pueda tener acceso al código fuente del sistema de IA y a los datos de entrenamiento para que puedan ser analizados por un experto. No debe permitirse que el secreto comercial afecte a los derechos de defensa.

32. Dado al elevado coste de obtener un análisis pericial de un sistema de IA, los Estados deben garantizar que cualquier persona acusada de un delito con base en una probabilidad calculada por un sistema de IA tenga acceso no sólo a una asistencia letrada efectiva, sino también a apoyo económico para dicha pericia especializada.

Sección IV: Perspectivas internacionales sobre la IA: Desafíos para la cooperación judicial y el Derecho Penal/Humanitario Internacional

Preámbulo

Considerando que la inteligencia artificial (IA) es parte integrante del proceso de toma de decisiones de los Sistemas de Armas Autónomos (SAA), y que tales armas ya han alcanzado un nivel significativo de desarrollo en esta tercera década del siglo XXI y que ya son ampliamente utilizadas por muchos Estados y actores no estatales;

Considerando que múltiples decisiones de vigilancia y selección de objetivos tradicionalmente adoptadas o informadas por seres humanos están empezando a automatizarse mediante el uso de los SAA, lo que provoca importantes consideraciones jurídicas, morales y éticas;

Destacando la creciente preocupación por la nocividad de los usos malintencionados o negligentes de los SAA;

Reconociendo la necesidad de analizar si la respuesta legal de los Estados y de la comunidad internacional a los retos de la IA es suficiente o si necesita ser reformada y adaptada, ya sea mediante modificaciones específicas o mediante la creación de nuevas formas de criminalización;

Reconociendo la necesidad de que los SAA se diseñen y desarrollen de conformidad con el derecho internacional.

* * * *

Observando que los SAA pueden definirse como '[c]ualquier sistema de armas con autonomía en sus funciones críticas, es decir, un sistema de armas que puede seleccionar (buscar, detectar, identificar, rastrear o seleccionar) y atacar (usar la fuerza contra, neutralizar, dañar o destruir) objetivos sin intervención humana;'[4]

Observando, sin embargo, que esta definición de SAA no es universalmente aceptada, y que diferentes regímenes internacionales, regionales y nacionales pueden optar por adoptar diferentes interpretaciones de la definición de SAA y del control humano significativo en su uso ('*human on the loop*');

Observando que diferentes regímenes internacionales, regionales y nacionales también pueden optar por distinguir entre los SAA y los Sistemas de Armas Autónomos Letales, y también entre los SAA y los Sistemas de Armas Automatizados;

Considerando que el uso de los SAA puede causar daños significativos a los seres humanos y a las sociedades, y que dicho uso plantea implicaciones jurídicas y éticas relacionadas tanto con el *ius ad bellum* como con el *ius in bello*, violaciones graves de los derechos humanos, violaciones del *ius cogens* y principales crímenes internacionales;

Considerando que el uso de los SAA puede influir en la opinión pública y en las políticas a favor del uso de la fuerza, debido a la percepción de que el uso de los SAA minimiza los riesgos de destrucción, muerte o lesiones corporales de los soldados y otras personas implicadas;

Considerando que el uso de los SAA puede afectar negativamente al respeto de los principios fundamentales del *ius in bello*, como los principios de distinción, proporcionalidad y precaución;

Considerando que, al eliminar o reducir el elemento humano en la toma de decisiones, el uso de los SAA puede contribuir a aumentar el número de muertes debido a la ausencia de sentimientos humanos, como el miedo y la compasión, que pueden contribuir a reducir el número de muertes;

Considerando que el uso de los SAA puede causar destrucción y daños colaterales significativos;

Considerando que los principales crímenes internacionales pueden cometerse mediante el uso de los SAA y que dicho uso de los SAA puede plantear cuestiones graves y nuevas relacionadas con la atribución de responsabilidad penal, incluidas, entre otras, cuestiones relacionadas con la responsabilidad de mando.

* * * *

[4] CICR, *Opiniones del CICR sobre los sistemas de armas autónomos*, documento presentado a la Reunión de Expertos sobre Sistemas de Armas Autónomos Letales (SAAL) de la Convención sobre Ciertas Armas Convencionales, 11 de abril de 2016, https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system

Considerando que el uso de los SAA puede plantear cuestiones jurisdiccionales, porque el uso de los SAA puede ser trans-territorial;

Constatando que es necesario un enfoque regulador internacional de los SAA, que debería incluir una norma jurídicamente vinculante que prohíba el diseño, desarrollo, producción y uso de los SAA totalmente autónomos (sin un control humano significativo);

Animando a todos los Estados a entablar negociaciones oficiales en el seno de las Naciones Unidas (bajo los auspicios de la Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales y/o en el seno de otras convenciones, órganos, organismos e instituciones apropiados de las Naciones Unidas) con el objetivo de desarrollar normativas u otras directrices que se apliquen al diseño, desarrollo, producción y uso de los SAA.

Concluyendo que esto también refuerza la necesidad de un enfoque global de los SAA.

* * * *

Teniendo en cuenta

- el *Informe de la reunión informal de expertos de 2016 sobre Sistemas de Armas Autónomos Letales (SAAL)* de Naciones Unidas;[5]

- la Convención de Naciones Unidas sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que Puedan Considerarse Excesivamente Nocivas o de Efectos Indiscriminados, reformada el 21 de diciembre de 2001;[6]

- la Resolución del Parlamento Europeo, de 12 de septiembre de 2018, sobre los Sistemas de Armas Autónomos;[7]

- el Comunicado de la Conferencia Latinoamericana y del Caribe sobre el Impacto Social y Humanitario de las Armas Autónomas, 23 y 24 de febrero de 2023;

- el Informe del Comité Internacional de la Cruz Roja, *Sistemas de Armas Autónomos: Implicaciones de la Creciente Autonomía en las Funciones Críticas de las Armas*, Ginebra, septiembre de 2016;[8]

---

[5] https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons__Informal_Meeting_of_Experts_ (2016)/ReportLAWS_2016_AdvancedVersion.pdf

[6] https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/#:~:text=It%20was%20adopted %20on2010,or%20to%20affect%20civilians%20indiscriminately

[7] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_ES.html

[8] CICR, *Sistemas de Armas Autónomos: Implicaciones de la Creciente Autonomía en las Funciones Críticas de las Armas*, Ginebra, septiembre de 2016, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems

- la *Declaración ante la Reunión de Expertos sobre Sistemas de Armas Autónomos Letales (SAAL) de la Convención sobre Ciertas Armas Convencionales*, 13-17 de abril de 2015, Comité Internacional de la Cruz Roja, Ginebra;[9]

- el XX Congreso de la AIDP, Justicia penal y actividad empresarial, Roma, 13-16 de noviembre de 2019.[10]

***

Los participantes en el Coloquio Internacional de la Sección IV, celebrado en Opatija, Croacia, del 7 al 8 de diciembre de 2023, han aprobado lo siguiente:

Resoluciones

1) Con el fin de prevenir y reducir los daños relacionados con las SAA, es necesario que los legisladores internacionales, regionales y nacionales, así como otras autoridades competentes, definan plenamente las SAA y elaboren normativas que regulen su diseño, desarrollo, producción y uso.

2) El uso de los SAA debe estar regulado de antemano por una ley o una norma de fuerza vinculante equivalente.[11]

3) Se insta a los Estados a ser más transparentes sobre su uso de los SAA.

4) Los Estados deben asegurarse de que las decisiones tomadas por las autoridades para utilizar SAA se basan en criterios transparentes que están sujetos al escrutinio público. Los Estados que utilicen SAA deben elegir sistemas cuyo funcionamiento sea totalmente transparente, explicable y rastreable (IA de caja blanca).

5) La ley u otras normas equivalentes relacionadas con el uso de los SAA deben exigir que dichos sistemas tengan un alto grado de fiabilidad tecnológica. Una regulación suficientemente precisa que exija evaluaciones y validaciones adecuadas, independientes del desarrollador del SAA, debe limitar en la mayor medida posible el riesgo de sesgo o cualquier forma de discriminación en el aprendizaje automático, los errores de codificación y otros fallos tecnológicos.

[9] CICR (2015), Declaración ante la Reunión de Expertos sobre Sistemas de Armas Autónomos Letales (SAAL) de la Convención sobre Ciertas Armas Convencionales, 13-17 de abril de 2015, Ginebra, https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS

[10] https://www.penal.org/fr/aidp-xx-international-congress-penal-law-%E2%80%9Ccriminal-justice-and-corporate-business%E2%80%9D-rome-13-16-0.

[11] El artículo 36 del Protocolo Adicional I a los Convenios de Ginebra ya exige a los Estados miembros que sigan evaluando si el desarrollo o el uso de cualquier nueva arma estaría prohibido por el Protocolo Adicional o por cualquier otra norma de derecho internacional. Protocolo adicional a los Convenios de Ginebra del 12 de agosto de 1949 relativo a la protección de las víctimas de los conflictos armados internacionales (Protocolo I), artículo 36, https://www.ohchr.org/es/instruments-mechanisms/instruments/protocol-additional-geneva-conventions-12-august-1949-and .

6) Las leyes o normas equivalentes deben exigir que, antes de que un SAA basado en algoritmos de autoaprendizaje pueda utilizarse para la vigilancia, la selección de objetivos o cualquier otro fin, los algoritmos se desarrollen, entrenen, prueben y utilicen bajo supervisión humana (principio human-in-the-loop). Estas leyes y normas equivalentes deben exigir una evaluación humana antes de emprender cualquier acción para utilizar un SAA.

7) Los Estados y las autoridades encargadas de hacer cumplir la ley deben velar por que el personal que utiliza los SAA reciba una formación adecuada sobre el uso correcto de dichos sistemas, así como sobre los riesgos de error y sesgo. Deben asegurarse de que dicho personal tenga un conocimiento profundo de los peligros que los SAA pueden suponer para los derechos humanos.

8) Las leyes o normas equivalentes deben establecer explícitamente que cuando se produzca una violación de los derechos humanos como consecuencia de un mal funcionamiento de un SAA, la empresa que diseñó o fabricó el sistema incurrirá en responsabilidad penal por culpa o negligencia,[12] sin excluir la responsabilidad civil por productos defectuosos. Dichas leyes o normas equivalentes también deben prever que se lleven a cabo investigaciones para determinar la causa de la violación.

9) Todos los Estados que decidan utilizar SAA deben disponer de leyes penales adecuadas que establezcan que las personas que utilicen SAA de manera incompatible con las normas jurídicas internacionales y nacionales pertinentes incurrirán en responsabilidad penal individual ante las jurisdicciones penales competentes.[13]

10) Los Estados y los organismos regionales e internacionales de derechos humanos deben garantizar que el uso de los SAA no da lugar a violaciones graves de los derechos humanos, violaciones del ius cogens u otras violaciones de las normas pertinentes del derecho internacional y del derecho internacional humanitario.

11) Los Estados deben garantizar que han implementado plenamente sus obligaciones jurídicas basadas en el derecho internacional en sus respectivos ordenamientos jurídicos nacionales, a través de leyes, reglamentos, órdenes ejecutivas, códigos militares u otros tipos de normas jurídicas nacionales vinculantes. Dicha legislación nacional debe prever diferentes modos de responsabilidad, incluida, entre otras, la responsabilidad de mando.

[12] XXI Congreso de la AIDP: Inteligencia Artificial y Justicia Penal, Sección I: Categorías de Derecho Penal Tradicional e IA, Resoluciones, aprobadas por el Coloquio de Siracusa, 15-17 de septiembre de 2022; XXI Congreso AIDP: Inteligencia Artificial y Justicia Penal, Sección II: Derecho Penal y Criminalización ante los Desafíos de la IA, Resoluciones, aprobadas por el Coloquio Internacional de Bucarest, 16-16 de junio de 2023.

[13] XXI Congreso de la AIDP: Inteligencia Artificial y Justicia Penal, Sección I: Categorías de Derecho Penal Tradicional e IA, Resoluciones, aprobadas por el Coloquio de Siracusa, 15-17 de septiembre de 2022.

12) Los Estados deben establecer órganos de investigación independientes que estudien, de forma continua, si el uso de los SAA se está llevando a cabo de forma coherente con todas las normas y reglamentos jurídicos nacionales e internacionales aplicables.

13) Es necesario identificar y definir modalidades específicas de atribución de responsabilidad penal a las personas (tanto físicas como jurídicas) responsables del desarrollo de los SAA.[14] Dicha responsabilidad no debe excluir a las personas, físicas o jurídicas, que contribuyen a la cadena causal del daño: desde la persona que lo diseña, programa, produce, vende, distribuye hasta las propias personas usuarias finales de los sistemas.

14) Los Estados deben esforzarse por fomentar la cooperación entre las personas del ámbito político y académico expertas en materia de derecho internacional y los miembros de las fuerzas militares, policiales o de inteligencia encargados de desarrollar y/o desplegar SAA. Dicha cooperación debe llevarse a cabo con el objetivo de garantizar que el uso de los SAA sea siempre coherente con las normas pertinentes del derecho internacional, así como con las leyes y reglamentos nacionales aplicables.

15) Los Estados deben prever la aplicación extraterritorial de su legislación nacional en los casos en que los Estados o las entidades bajo control estatal desplieguen o utilicen de otro modo SAA en el extranjero.

16) Los Estados deben elaborar leyes adecuadas sobre conflictos de leyes y conflictos de jurisdicciones que permitan la aplicación de la legislación nacional en los casos en que los SAA se desplieguen o utilicen de otro modo en el extranjero. Dichas leyes pueden elaborarse a nivel nacional, regional y/o internacional. Dichas leyes deben permitir la imposición de responsabilidad penal (así como responsabilidad civil) a aquellas personas físicas o jurídicas bajo el control de los Estados pertinentes que sean responsables del mal uso o mal funcionamiento de los SAA ubicados fuera del territorio del Estado pertinente.

17) Los Estados deben desarrollar modelos de cooperación judicial en materia penal, especialmente para recabar pruebas relacionadas con delitos causados por los SAA. Dichos modelos de cooperación judicial pueden ser especialmente necesarios para los Estados que aún no hayan celebrado acuerdos bilaterales, regionales o internacionales adecuados en esta materia.

18) Los Estados deben asegurarse de que disponen de mecanismos de extradición para perseguir y castigar eficazmente a los responsables de delitos causados por los SAA.

[14] XXI Congreso de la AIDP: Inteligencia Artificial y Justicia Penal, Sección I: Categorías de Derecho Penal Tradicional e IA, Resoluciones, aprobadas por el Coloquio de Siracusa, 15-17 de septiembre de 2022.

Subscriptions & Membership Applications

AIDP/IAPL Membership

Annual Contribution of € 110

Benefactor Member – A member wishing to provide extra financial support to the Association. This type of membership includes subscription to the RIDP as well as online access.

Collective Member – Universities, associations, institutes, etc. This type of membership includes subscription to the RIDP.

National group – AIDP Members have established in numerous countries a National Group, which carries out its own scientific activities. Each National Group, in addition to the fees for individual members, has to pay to the AIDP a membership fee which entitles the national group to participate in the activities of the Association. This type of membership includes subscription to the RIDP.

Annual Contribution of € 85

Individual Member – The AIDP membership includes subscription the RIDP as well as online access to the RIDP archives and the RIDP *libri* series.

Annual Contribution of € 45

Young Penalist – AIDP members under the age of 35 may join the Young Penalist Group, which carries out its own activities and elects representatives to the organs of AIDP. This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Student or retiree – AIDP membership for a reduced contribution. This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Reduced-fee countries – If you are residing in a country listed on the reduced country fee list, you will be entitled to membership including a subscription to the RIDP for a limited membership fee. The list can be consulted on the AIDP website under the section 'About Us' – guidelines to establish a national group. http://www.penal.org/en/guidelines-establishment-national-groups . This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Annual Contribution of € 40

AIDP Individual Membership without RIDP subscription - mere AIDP membership without RIDP subscription and no access to electronic archives.

Membership Application instructions

The membership application form can be downloaded at the AIDP website (http://www.penal.org/en/user/register) and returned by email or mail to the address below:

Email: secretariat@penal.org. Secretariat: AIDP, c/o The Siracusa International Institute, Via Logoteta 27, 96100 Siracusa, Italy

BNP PARIBAS Bordeaux C Rouge N° IBAN : FR76 3000 4003 2000 0104 3882 870

Payment instructions

By check: Join your check to your membership application form and mail it to: AIDP secretariat, c/o The Siracusa International Institute, Via Logoteta 27, 96100 Siracusa, Italy.

Bank transfer: The bank and account details are on the membership application form. Once the bank transfer is done, send your membership application form together with a copy of the bank transfer order by fax or email, or by mail to the secretariat of the Association. The identity of the sender does not appear on the bank statement and if you do not send a copy of the bank transfer separately, we will not be able to credit the transfer to your membership.

Payment by credit card: The cryptogram is the three-digit number on the reverse side of your credit card. It is necessary for payment. Do not forget to sign your application. Please return the form by fax or email, or by mail.

For further information please consult the AIDP website http://www.penal.org/.

Subscription to the RIDP

Single Issue – price indicated for each issue on MAKLU website.

Annual Subscription – For the price of € 85, an annual subscription to the RIDP can be obtained which includes the print and free online access to the RIDP back issues. This subscription does not include AIDP Membership.

For RIDP subscription, please follow the instructions on the MAKLU publisher's website: http://www.maklu-online.eu

Recognizing the profound impact of AI on criminal justice, the IAPL devoted its 2020-2024 scientific cycle and its concluding XXIst Congress to examining the transformative effects of AI and the legal challenges it poses for both substantive and procedural criminal law. The topic was prepared through a concept paper and explored through four international colloquia, each culminating in the adoption of a set of resolutions addressing a distinct aspect of the criminal justice system. This issue wraps up the cycle, presenting the concept paper (in English) and the trilingual version of the four sets of resolutions.

Au vu de l'impact profond de l'IA sur la justice pénale, l'AIDP a consacré son cycle scientifique 2020-2024 et son XXIe Congrès à l'étude des effets transformateurs de l'IA et des défis juridiques qu'elle pose dans le cadre du droit pénal et de la procédure pénale. Le sujet a été abordé grâce à un document de réflexion initiale et exploré par la suite lors de quatre colloques internationaux, chacun aboutissant à l'adoption d'une série de résolutions portant sur un aspect distinct du système de justice pénale. Ce numéro conclut ce cycle scientifique par la publication du document de réflexion (en anglais) et la version trilingue des quatre séries de résolutions.

Dado el profundo impacto de la IA en la justicia penal, la AIDP dedicó su ciclo científico 2020-2024 y su XXIo Congreso a examinar los efectos transformadores de la IA y los retos jurídicos que plantea tanto para el derecho penal sustantivo como para el procesal. El tema se inició con un documento conceptual y se exploró a través de cuatro coloquios internacionales, cada uno de los cuales culminó con la adopción de un conjunto de resoluciones que abordaban un aspecto distinto del sistema de justicia penal. Este número cierra este ciclo, presentando el documento conceptual (en inglés) y la versión trilingüe de los cuatro conjuntos de resoluciones.

Katalin Ligeti is President of the AIDP/IAPL and Dean of the Faculty of Law, Economics and Finance and Professor of European and International Criminal Law at the University of Luxembourg.

John A.E. Vervaele is Honorary President of the AIDP/IAPL, Emeritus Professor at Utrecht University, The Netherlands, and Professor in European Criminal Law and Human Rights at the College of Europe, Bruges, Belgium

Gert Vermeulen is General Director Publications of the AIDP/IAPL, Editor-in-chief of the RIDP, and Senior Full Professor of European and international criminal law, sexual criminal law, and data protection law at Ghent University, Belgium.

www.maklu.be
ISBN 978-90-466-1296-5